Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Ciência da Computação

Sumarização Abstrativa de Documentos Longos Utilizados em Fiscalizações e Instruções Processuais

Eric Hans Messias da Silva

Dissertação apresentada como requisito parcial para conclusão do
Mestrado Profissional em Computação Aplicada

Orientador
Prof. Dr. Marcelo Ladeira

Brasília
2023

Ficha catalográfica elaborada automaticamente,
com os dados fornecidos pelo(a) autor(a)

SS586s

Silva, Eric Hans
Sumarização Abstrativa de Documentos Longos Utilizados em Fiscalizações e Instruções Processuais / Eric Hans Silva; orientador Marcelo Ladeira. -- Brasília, 2023.
106 p.

Dissertação(Mestrado Profissional em Computação Aplicada) -- Universidade de Brasília, 2023.

1. Inteligência Artificial. 2. Ciência de Dados. 3. Processamento de Linguagem Natural. 4. Sumarização Abstrativa de Textos Longos com conteúdo jurídico. 5. Métricas de sumarização. I. Ladeira, Marcelo, orient. II. Título.

Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Ciência da Computação

Sumarização Abstrativa de Documentos Longos Utilizados em Fiscalizações e Instruções Processuais

Eric Hans Messias da Silva

Dissertação apresentada como requisito parcial para conclusão do
Mestrado Profissional em Computação Aplicada

Prof. Dr. Marcelo Ladeira (Orientador)
CIC/UnB

Prof. Dr. Andrei Lima Queiroz
Universidade de Brasília

Prof. Dr. Thiago de Paulo Faleiros
Universidade de Brasília

Dr. Thiago Alexandre Salgueiro Pardo
Universidade de São Paulo

Prof. Dr. Gladston Luiz da Silva
Coordenador do Programa de Pós-graduação em Computação Aplicada

Brasília, 13 de julho de 2023

Dedicatória

Dedico este trabalho, em primeiro lugar, à minha família: aos meus pais, Lúcio e Waldja, e à minha irmã, Nathasha. Vocês constituem a base da primeira família que temos, responsável por moldar quem somos e nos preparar para viver neste mundo repleto de oportunidades e, principalmente, desafios.

Dedico também este trabalho à minha esposa, Ronia, que tem estado ao meu lado por quase 20 anos, sempre me motivando e apoiando em cada desafio que enfrento. Dedico também este trabalho aos nossos filhos, Gabriel e Davi, que estão começando a caminhar nessa jornada de estudos. Espero ser uma inspiração para o futuro deles, para que se desenvolvam e se aperfeiçoem em todos os aspectos de suas vidas. Todos vocês são extremamente importantes para mim e, sem vocês, eu não teria a energia para ir ainda mais longe.

Agradecimentos

Agradeço ao Tribunal de Contas da União por ter me concedido preciosas horas para o desenvolvimento deste trabalho, além de fornecer insumos valiosos para a minha pesquisa. Foi um trabalho desafiador e gratificante poder retribuir as horas concedidas com um produto que pode ser utilizado para aprimorar o trabalho dos nossos servidores.

Expresso minha gratidão ao meu orientador, Professor Marcelo Ladeira, por ter me orientado nesse processo de desenvolvimento de um trabalho na área de Inteligência Artificial. Esta área, em constante e radical evolução, apresenta mudanças significativas que impactam a maneira como a enxergamos a cada dia. Não foi uma tarefa simples, demandando muito tempo para manter o trabalho relevante e atualizado. Muito obrigado!

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES), por meio do Acesso ao Portal de Periódicos.

Resumo

O Tribunal de Contas da União tem seu trabalho organizado por processos e, ao longo do ciclo de vida deles, cada processo chega geralmente a conter de dezenas a centenas de peças processuais. Cada peça atinge facilmente algumas dezenas de páginas. A quantidade de processos e documentos só tende a crescer ao longo do tempo, o que gera uma quantidade enorme de material para leitura e com conteúdo bem rico, mas de difícil consumo, pois é necessário um tempo considerável para a leitura de cada processo. Os processos costumam ser lidos para verificar se possuem conteúdo relevante para alguma fiscalização ou instrução processual em curso. Além do custo alto para ler um processo, parte desse conteúdo é descartado pelo auditor por não estar atrelado ao seu trabalho corrente, o que gera um desperdício de tempo nesta atividade. Para melhorar a eficiência deste processo, é proposto neste trabalho o desenvolvimento de uma solução de sumarização automática de texto usando aprendizado de máquina aplicado ao processamento de linguagem natural. Essa solução utiliza a abordagem de sumarização híbrida (extrativa combinada com abstrativa) aplicada a documentos longos e com conteúdo jurídico. A solução foi disponibilizada como uma aplicação Web com microsserviço para melhor integração com aplicações que compõem o processo de trabalho do auditor. Os resumos gerados pelos modelos foram avaliados principalmente por métricas que foquem mais na semântica do texto gerado e, em decorrência disso, têm uma melhor aderência ao conteúdo desejado.

Palavras-chave: Processamento de Linguagem Natural, Sumarização Abstrativa, Documentos Longos, Documentos Jurídicos

Abstract

The Brazilian Federal Court of Accounts organizes its work by processes and, throughout their life cycle, each of them usually contains from tens to hundreds of legal documents. Each document easily reaches a few dozen pages. The number of processes and documents only tends to grow over time, which generates a huge amount of material for reading and with a very rich content, but difficult to consume, as it takes considerable time to read each process. The processes are usually read to verify if they have relevant content for any fiscalization or procedural instruction in progress. In addition to the high cost of reading a process, part of this content is discarded by the auditor because it is not linked to their current work, which generates a waste of time in this activity. To improve the efficiency of this process, we proposed in this work the development of an automatic text summarization solution using machine learning applied to natural language processing. This solution uses the hybrid summarization approach (extractive combined with abstractive) applied to long documents with legal content. The solution was made available as a Web application with microservice for better integration with applications that make up the auditor's work process. The summaries generated by the models were evaluated mainly by metrics that focus more on the semantics of the generated text and, as a result, have better adherence to the desired content.

Keywords: Natural Language Processing, Abstractive Summarization, Long Documents, Legal Documents

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas e Siglas

BERT *Bidirectional Encoder Representations from Transformers.*

DUC *Document Understanding Conferences.*

EHE *Extraction History Encoder.*

GCE *Global Context Encoder.*

GPU *Graphics Processing Unit.*

GRU *Gated Recurrent Unit.*

IDF *Inverse Document Frequency.*

LCS *Longest Common Subsequence.*

LED *Longformer-Encoder-Decoder.*

LSE *Local Sentence Encoder.*

LSTM *Long Short-Term Memory.*

MHA *Multi-Head self-Attention.*

MHP *Multi-Head Pooling.*

MLM *Masked Language Modeling.*

NMT *Neural Machine Translation.*

NSP *Next Sentence Prediction.*

PLN Processamento de Linguagem Natural.

REST *Representational State Transfer.*

RLHF *Reinforcement Learning from Human Feedback.*

RNN *Recurrent Neural Network.*

Rouge *Recall-Oriented Understudy for Gisting Evaluation.*

Seq2Seq *Sequence to Sequence.*

STF Supremo Tribunal Federal.

TAC *Text Analysis Conference.*

TCU Tribunal de Contas da União.

TF-IDF *Term Frequency–Inverse Document Frequency.*

TPU *Tensor Processing Units.*

WMD *Word Mover's Distance.*

Capítulo 1

Introdução

As funções básicas do Tribunal de Contas da União (TCU) podem ser agrupadas da seguinte forma: fiscalizadora, consultiva, informativa, judicante, sancionadora, corretiva, normativa e de ouvidoria. Algumas de suas atuações assumem ainda o caráter educativo [1]. Há cinco instrumentos por meio dos quais se realiza a fiscalização: levantamento, auditoria, inspeção, acompanhamento e monitoramento [1]. As prestações de contas, as fiscalizações e demais assuntos submetidos à deliberação do Tribunal organizam-se em processos [1].

Os processos vão acumulando peças ao longo de sua vida e chegam geralmente a conter de dezenas a centenas delas e podem chegar a milhares em alguns casos. Cada documento chega comumente a algumas dezenas de páginas. O volume de processos e peças só tende a crescer ao longo do tempo.

A informação contida nesses documentos é bastante rica, pois descreve tudo o que ocorreu dentro do processo nos mínimos detalhes e fica acessível para que os auditores do TCU possam ler posteriormente, mas o uso efetivo dessas informações é um desafio até hoje no TCU, pois a quantidade de informação não estruturada é muito grande e de difícil consulta rápida.

Atualmente, para consumir as referidas informações, os auditores necessitam inicialmente pesquisar por palavras-chaves em peças processuais para, em seguida, ler seu conteúdo e avaliar sua pertinência ao tema desejado. Esse é um ciclo que se repete para praticamente todos os auditores: filtrar peças por palavras-chaves, ler seu conteúdo, selecionar as mais importantes e voltar para o primeiro passo para refinamento ou para procurar por outro assunto com outro conjunto de palavras-chaves.

Essa atividade consome bastante tempo, pois além de refinar pesquisas por palavras-chaves, é necessário em muitos casos ler grande parte dos documentos, que são longos, para avaliar se seu conteúdo é útil para subsidiar alguma decisão sobre investigações em curso.

O Uso de técnicas aprendizado de máquina aplicadas a Processamento de Linguagem Natural (PLN) tem trazido resultados satisfatórios para obtenção de informação baseada em texto com um grau de sofisticação maior do que uma pesquisa textual por palavra-chave poderia trazer.

Dentre as técnicas de aprendizado de máquina, uma que pode contribuir para otimização da leitura de peças processuais a fim de obter informações de maneira mais rápida é o uso de geração de resumos (sumarização), pois pode-se gerar um pequeno extrato para cada documento necessário em uma fiscalização e obter as partes mais relevantes para o auditor. Com isso, ele poderá avaliar rapidamente se o documento é relevante e requer a leitura completa ou descartá-lo, sem a necessidade da leitura de boa parte dele, caso não seja relacionado à investigação em curso durante a fiscalização.

1.1 Definição do Problema

Os processos do TCU compõem uma grande fonte de conhecimento do que é produzido no Tribunal, pois todas suas decisões estão contidas em suas peças. Processos ativos ou encerrados são utilizados como fonte de conhecimento, pois são consultados com frequência sobre temas tratados por eles, com o objetivo de analisar decisões anteriores bem como verificar o que já existe firmado de precedente ou jurisprudência que possa ser aplicado ao que se está investigando no momento.

Cada processo possui uma variedade grande de peças com diferentes tamanhos e estruturas. Alguns dos documentos são curtos e não necessitam de sumarização, como extratos bancários e avisos de recebimentos de correspondências, mas os documentos mais relevantes e que tratam do cerne do processo são mais longos, como: Instruções Processuais, Relatórios, Votos e Acórdãos.

Os auditores do TCU apontaram, em um questionário realizado internamente (Apêndice A), que os seguintes documentos seriam os mais beneficiados com a geração automática de resumos (múltiplas respostas possíveis):

- **Instrução Processual**: apontado por 63,9% dos auditores,
- **Relatório**: apontado por 58,3% dos auditores,
- **Resposta à Comunicação**: apontado por 44,4% dos auditores,
- **Acórdão**: apontado por 25% dos auditores,
- **Voto**: apontado por 22,2% dos auditores,
- **Normativo**: apontado por 2,8% dos auditores.

Os dois primeiros documentos, apesar de não serem lidos mais frequentemente que os acórdãos, são, em geral, muito mais longos, tornando-os candidatos naturais para a sumarização. Para uma delimitação mais eficaz do escopo, o documento de Instrução Processual foi selecionado como objeto deste trabalho.

A tarefa de sumarizar a Instrução Processual representa um considerável desafio na área de aprendizado de máquina, uma vez que este tipo de documento pode ultrapassar a marca de 100 páginas em determinados processos. No contexto desta pesquisa, foi examinada a possibilidade de se realizar o resumo de apenas uma seção do documento em questão, mais precisamente, o Exame Técnico, seção esta que foi considerada como a mais relevante pelos auditores no questionário apresentado no Apêndice A. O Exame Técnico, por sua vez, possui extensão significativa, podendo ser categorizado como um documento longo por si só, ultrapassando, em termos de tamanho, o que usualmente se emprega em *datasets* de treinamento para este tipo de tarefa.

A sumarização de texto, quanto à sua geração, pode ser classificada em extrativa, abstrativa ou híbrida. A extrativa, em geral, tem uma etapa de seleção de partes relevantes do texto para posterior combinação dessas sentenças para a geração do resumo. A abstrativa, por outro lado, processa o texto por inteiro e o reesecreve como uma paráfrase comprimida do texto original. A híbrida combina a extrativa e abstrativa com o objetivo de reduzir seus problemas e melhorar seus resultados [2].

Neste trabalho, os experimentos com sumarização abstrativa utilizam apenas a seção de Exame Técnico enquanto que os das sumarizações extrativas e híbridas utilizam com o documento com seu inteiro teor. Mais detalhes sobre as motivações dessas variações podem ser encontrados no Capítulo 4.

Cada uma das abordagens para geração de resumo traz consigo alguns problemas:

- **Sumarização Extrativa**: a seleção de sentenças mais importantes faz com que o resumo contenha textos importantes, porém desconexos entre si, o que dificulta o leitor a entender o contexto geral do texto original.
- **Sumarização Abstrativa**: problemas de alucinação em que o resumo gerado pode contradizer (alucinação intrínseca) o texto original ou mesmo extrapolar e adicionar ideais que não encontram suporte no texto original (alucinação extrínseca) [3].
- **Sumarização Híbrida**: este tipo de geração quando combina os dois formatos acima pode ter acúmulo dos erros inerentes a cada abordagem.

Além dos problemas citados acima para cada abordagem de geração de resumo, a sumarização de documentos longos torna o problema mais complexo.

Notícias de jornal, que são bastante usadas na literatura para resumos (não longos), ficam em torno de 400 a 800 palavras de texto para resumir [4, 5]. Isso equivale, no pior

caso (800 palavras) a algo em torno de 1 página cheia. Neste trabalho, busca-se resumir documentos que ficam entre 20 a 30 páginas, o que representa um aumento significativo da entrada e uma dificuldade maior em se extrair do texto o que é mais relevante para o auditor.

Outro problema relacionado à sumarização em geral é como avaliar bem seu resultado de maneira automatizada e baseado na sua semântica, pois tarefas de geração de linguagem natural aceitam muitas variações sem que seu significado seja prejudicado.

Outro problema relacionado à sumarização em geral é como avaliar bem seu resultado de maneira automatizada e baseado na semântica do texto gerado, pois tarefas de geração de linguagem natural podem ter como saída inúmeras variações sem que o significado do texto gerado seja prejudicado.

Algumas questões que serão objetos desta pesquisa são:

1. Como propor uma abordagem de sumarização abstrativa aplicada a documentos longos em português brasileiro com conteúdo jurídico que possa facilitar o trabalho do auditor no TCU na seleção de instruções processuais para leitura?
2. Qual métrica automatizada possui melhor aderência com a avaliação do auditor para o conteúdo do resumo gerado?

1.2 Justificativa do Tema

A leitura de documentos extensos em fiscalizações ou instruções processuais é uma atividade bastante recorrente e de execução lenta por ter que se debruçar sobre uma quantidade muito grande de documentos apenas para avaliar se seu conteúdo é pertinente ao trabalho sendo realizado. Muitos documentos são descartados nesta fase, o que gera um desperdício de tempo.

Além disso, a sumarização de peças processuais do próprio TCU é útil para obter informações rápidas sobre pessoas, empresas e órgãos a partir de diferentes processos e ter uma visão global composta por múltiplos processos acerca de um determinado ente e, com isso, usar essa informação para priorização de futuras fiscalizações.

É possível montar uma base de conhecimento a partir desses resumos com informações sobre cada pessoa, seja física ou jurídica, e com isso, fazer análise de grafos procurando por ligações a partir de processos, bem como também o uso de *Question Answering* sobre os resumos, tornando o processo de obtenção de informação muito mais intuitivo e de fácil interação por parte do auditor do TCU.

A sumarização abstrativa foi escolhida como abordagem inicial, pois ela produz um texto mais próximo daquele gerado por seres humanos ao realizar a mesma tarefa. No

início deste trabalho, início de 2022, os modelos que apresentavam melhores desempenhos eram aqueles que empregavam abordagens abstrativas ou híbridas baseadas em Modelos de Linguagem. Contudo, com a emergência dos Modelos de Linguagem Grandes ao final de 2022 e início de 2023 e seu fácil acesso, houve uma necessidade de reavaliar e evoluir as abordagens, de forma a considerar esses novos modelos.

O objetivo de obter resumo é permitir que o leitor compreenda rapidamente o conteúdo do documento, avaliando sua relevância e decidindo se vale a pena a leitura completa ou se deveria ser rapidamente descartado. Não se pressupõe utilizar o resumo como substituto do texto original.

A plataforma construída para a solução do problema se propõe a ser um ponto central de investigação de pessoas físicas e jurídicas durante fiscalizações e instruções processuais por parte dos auditores do TCU.

1.3 Objetivos

Abaixo são listados os objetivos gerais e específicos de forma a desenvolver uma solução de sumarização automática de textos disponibilizada como um microsserviço de forma integrada ao processo de trabalho do auditor do TCU.

1.3.1 Objetivo Geral

O objetivo geral deste projeto é gerar resumos de textos de documentos longos em português com conteúdo jurídico, utilizados em fiscalizações ou instruções processuais no TCU para facilitar a seleção de textos para leitura, durante o trabalho dos auditores.

Para melhor avaliação do modelo, será escolhida uma métrica que tenha mais alta aderência com a avaliação do auditor para que os modelos possam ser avaliados de maneira mais próxima de uma avaliação humana, porém de forma automatizada.

1.3.2 Objetivos Específicos

- Pré-treinar Modelos de Linguagem baseados no *Longformer-Encoder-Decoder* (LED) [6] com textos em português.
- Pré-treinar Modelos de Linguagem baseados no LED [6] com textos em português com conteúdo jurídico.
- Realizar ajuste fino de modelos baseados no LED [6] para sumarização de documentos longos em português.

- Realizar ajuste fino de modelos baseados no LED [6] para sumarização de documentos longos em português com conteúdo jurídico.
- Comparar métricas de sumarização para escolha da que tiver a mais alta correlação com a avaliação humana, realizada pelo auditor, dos resumos gerados.
- Comparar modelos gerados para escolha do melhor sumarizador utilizando a métrica escolhida.
- Desenvolver uma aplicação Web com serviço REST [7] para expor um *endpoint* para receber um texto em formato PDF para extrair o texto para sumarização.
- Integração do serviço com fluxo de trabalho do auditor no sistema de processos do TCU e disponibilização de ferramenta para colher feedback do usuário sobre o resumo gerado.

1.4 Hipóteses de Pesquisa

- Modelos de aprendizado de máquina baseados em *Transformers* como o *Longformer-Encoder-Decoder* (LED) e *Big Bird-Pegasus* apresentam os melhores resultados em sumarização abstrativa aplicada a documentos longos e podem ser pré-treinados e realizado ajuste fino em textos em português com conteúdo e vocabulário jurídico para melhorar o desempenho do modelo na sumarização das peças de Instrução Processual no TCU.
- As métricas semânticas como BERTScore e MoverScore avaliam o conteúdo do documento ao invés de posições e ordem de palavras em textos, como o Rouge, e se adaptam melhor a variações de texto que não prejudiquem o significado geral do resumo.

1.5 Contribuições Esperadas

Este trabalho contribui com o desenvolvimento de uma solução de geração automática de resumos abstrativos aplicados a textos em português e com conteúdo jurídico. A geração desses resumos vai permitir ao auditor conhecer os fatos mais relevantes constantes nos documentos e, com isso, evitará que ele tenha que ler seu inteiro teor, o que além de consumir tempo, pode distraí-lo de fatos relevantes para o processo. A solução será integrada ao processo de trabalho do auditor e com possibilidade de feedback, para posterior análise dos resumos com o intuito de haver uma melhoria contínua de seus resultados.

Serão disponibilizados para a comunidade ao fim deste trabalho:

- Modelos de Linguagem treinados para o português
- Modelos de Linguagem treinados para o português jurídico
- Modelos de sumarização de textos jurídicos do TCU
- Código-fonte usado para pré-treino em português
- Código-fonte usado para pré-treino em português jurídico
- Código-fonte usado para ajuste fino em sumarização

1.6 Organização do Trabalho

Este documento está organizado da seguinte forma:

- Capítulo 2: neste capítulo é apresentada uma revisão do estado da arte, e os trabalhos relevantes na área de sumarização automática de textos que serviram de inspiração para o trabalho corrente, como os modelos que são pilares atuais para o estado da arte em sumarização: BART [8] e *Pegasus* [9], bem como modelos construídos a partir deles com foco em documentos longos como o LED [6] e o *Big Bird-Pegasus* [10].
- Capítulo 3: neste capítulo são detalhados aspectos que embasam todo o trabalho, desde a base utilizada pelos modelos, os *Transformers* [11], passando pelos mecanismos de atenção [6, 10] propostos para tratar documentos mais longos até as métricas que são usadas para avaliar o modelo, como Rouge-1/-2/-L [12], MoverScore [13] e BERTScore [14]. Além disso, serão detalhados os modelos mais recentes empregados neste trabalho, como o ChatGPT [15] e o MemSum [16], que compuseram a solução final deste trabalho.
- Capítulo 4: este capítulo detalha de forma minuciosa todos os desafios enfrentados na proposição da solução. São detalhados desde os *datasets* utilizados ou criados no escopo deste trabalho até todas as variações de modelos de sumarização extrativa, abstrativa e híbrida. Por fim, é detalhada a metodologia de avaliação humana empregada e sua comparação com as métricas automatizadas escolhidas para avaliação de resumos neste trabalho.
- Capítulo 5: este capítulo traz as conclusões deste trabalho, respondendo as questões de pesquisa, apontando o modelo escolhido para sumarização e a métrica automatizada que tem melhor aderência com a avaliação humana feita pelo auditor. Além disso, aponta limitações da solução corrente, bem como trabalhos futuros para aprimorá-la.

Capítulo 2

Revisão do Estado da Arte

Sumarização automática de texto é a tarefa de Processamento de Linguagem Natural (PLN), responsável por reduzir o conteúdo do texto enquanto preserva seu significado original [17]. Essa tarefa pode ser dividida quanto a sua forma de geração de texto em extrativa, abstrativa ou híbrida. Na sumarização extrativa, em geral, o resumo concatena as sentenças mais relevantes do documento original. Na sumarização abstrativa, que é mais complexa, o texto é processado por inteiro e é gerado um novo texto menor com o mesmo sentido que o documento original [2]. Na sumarização híbrida, ocorre a combinação da técnica extrativa com a abstrativa. Muitos pesquisadores têm obtido resultados mais satisfatórios utilizando essa técnica em comparação com a abordagem apenas abstrativa ou com a utilização de técnicas de aprendizado por reforço [2]. A sumarização extrativa foi o foco primário de pesquisa, mas seu ritmo de desenvolvimento foi reduzindo ao longo dos anos e apenas melhorias marginais estão ocorrendo, enquanto que a sumarização abstrativa tem tido maior foco em pesquisa nos últimos anos em comparação com a extrativa [18].

O estado da arte na área de sumarização abstrativa tem sido dominado pelos modelos *Sequence to Sequence* (Seq2Seq), que são usados para mapear uma sequência de texto de entrada (denominada *encoder*) para outra sequência de texto de saída (denominada *decoder*). O Seq2Seq se baseia em *Recurrent Neural Network* (RNN) que utilizam *Long Short-Term Memory* (LSTM) ou *Gated Recurrent Unit* (GRU). Os componentes principais deste modelo, juntamente com suas respectivas funções, são:

- *Embeddings*: consistem na representação vetorial numérica dos textos que são manipulados pelo modelo;
- *Encoder*: atua como a entrada do modelo, recebendo os *embeddings* e usando RNNs;
- *Attention*: é responsável por realizar o alinhamento dos *tokens* de entrada do *encoder* com os *tokens* de saída do *decoder*;

- *Decoder*: recebe a saída do *encoder*, que é combinada com a *attention*, e gera a saída do modelo usando RNNs.

A estrutura do Seq2Seq é ilustrada na Figura 2.1.

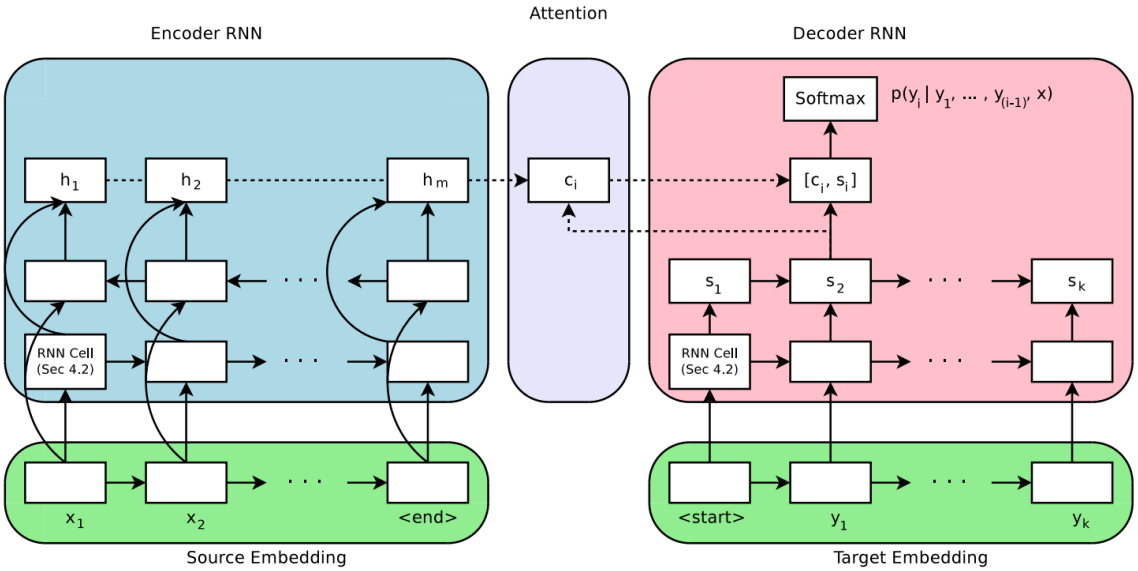

Figura 2.1: Modelo Seq2Seq baseado em RNN com mecanismo de atenção (Fonte: [19]).

Uma forte limitação dos modelos baseados em RNN, por causa da sua natureza sequencial, é que eles não podem tirar proveito de *Graphics Processing Unit* (GPU) ou *Tensor Processing Units* (TPU), capazes de paralelizar um volume massivo de operações. Além disso, esses modelos têm dificuldade de lidar com documentos muito longos, resultando em resumos de baixa qualidade [2].

Abordagens baseadas em aprendizado por reforço foram usadas para resolver algumas das limitações dos modelos Seq2Seq. Uma das vantagens dessas abordagens era usar a métrica desejada como parte da função de recompensa para otimização, ao invés da entropia cruzada como função *loss* na abordagem supervisionada, que não tem relação direta com o objetivo final. Entretanto as abordagens de aprendizado por reforço sozinhas ainda têm problema de generalização, que pode ser resolvido com o uso de *Transfer Learning* [2].

Nos últimos anos, os modelos que dominam o estado da arte em várias tarefas de PLN são baseados em *Transformers* [11], que eliminam a recorrência presente nos modelos Seq2Seq e são, portanto, computacionalmente mais eficientes. A Figura 2.2 mostra a arquitetura proposta por Vaswani et al. [11], onde tem-se no retângulo cinza à esquerda a representação dos N *encoders* enquanto que à direita tem-se o retângulo cinza, um pouco maior, a representação dos N *decoders*, como nos modelos Seq2Seq. O mecanismo de *self-attention*, nos *Transformers*, tem seu uso mais amplo e, diferentemente do Seq2Seq, está

embutido tanto nos *encoders* quanto nos *decoders*. O *self-attention* substitui completamente a necessidade recorrência e está representado pelos retângulos laranjas intitulados *Multi-Head Attention*.

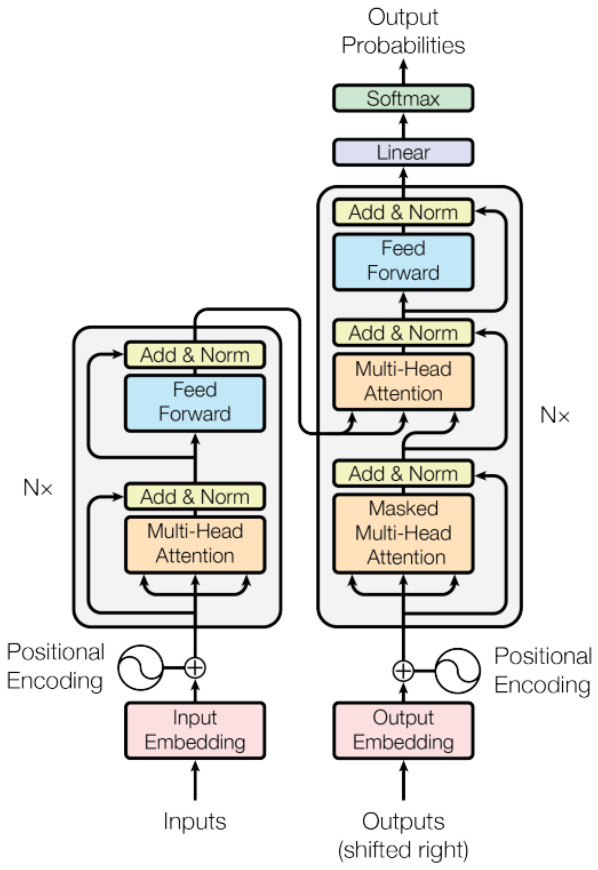

Figura 2.2: Arquitetura Transformer (Fonte: [11]).

Com o surgimento dos *Transformers* [11], vários modelos foram pré-treinados como Modelos de Linguagem, como é o caso do BERT [20], que estabeleceu o estado da arte simultaneamente em várias tarefas de PLN[1] a partir do mesmo pré-treino, apenas com ajuste fino para as tarefas específicas. Para sumarização, os modelos que ganharam mais destaque recentemente e estabeleceram o estado da arte em *datasets* destas tarefas foram o BART [8] e o *Pegasus* [9] ou os que os utilizaram como base para obter melhores resultados [2].

Os documentos de interesse deste trabalho são longos e representam um grande desafio para sumarização, pois eles possuem uma estrutura mais rica e dependem de mais contextos espalhados pelo texto [10]. Apesar do BART [8] e o *Pegasus* [9] terem estabelecido o estado da arte em sumarização, eles não conseguem processar documentos longos, pois

[1] `https://gluebenchmark.com/leaderboard`

possuem uma limitação de entrada de 1.024 *tokens* em virtude do mecanismo de atenção proposto por Vaswani et al. [11] crescer quadraticamente em função do tamanho da entrada, o que dificulta usá-los para os documentos do TCU.

Novos modelos foram propostos otimizando o mecanismo de atenção original, reduzindo sua complexidade de quadrática para linear, como o Longformer [6] e o *Big Bird* [10] para que consigam lidar melhor com documentos longos. Ambos os modelos são baseados no RoBERTa [21] e otimizam seu mecanismo de atenção. O Longformer [6] possui uma variação de modelo para tarefas de geração de linguagem natural, o *Longformer-Encoder-Decoder* (LED) [6], que otimiza o mecanismo de atenção do BART [8], enquanto que o *Big Bird* [10] tem uma variação na mesma linha só que ao invés do BART [8], é otimizado o mecanismo de atenção do *Pegasus* [9]. O LED, com isso, consegue estender o tamanho da entrada do BART [8] de 1.024 *tokens* para 16.384 *tokens*, enquanto que o *Big Bird* [10] consegue estender o tamanho da entrada do *Pegasus* [9] de 1.024 *tokens* para 4.096 *tokens*.

O LED [6] e o *Big Bird* [10] foram pré-treinados e tiveram o ajuste fino realizados em inglês, mas o conteúdo que deseja sumarizar é em português brasileiro com conteúdo jurídico. Para realizar o ajuste fino dos Modelos de Linguagem com melhores resultados, é necessário ter um *dataset* em português brasileiro, porém poucos trabalhos estão disponíveis aplicados ao nosso idioma. Entretanto, já existem alguns *datasets* para este fim, bem como modelos de linguagem pré-treinados em português usando como base o modelo T5 [22, 23].

Alguns modelos começaram a surgir para sumarização de documentos longos em português nos últimos anos. O PLSum [22] usou uma abordagem híbrida combinando o *Term Frequency–Inverse Document Frequency* (TF-IDF), para o primeiro estágio de sumarização extrativa, com os modelos baseados em *Transformers* para a parte abstrativa. Os autores experimentaram com dois modelos neste estágio: O PTT5 [24], modelo pré-treinado para o português e o *Longformer-Encoder-Decoder* (LED) [6] original, sem pré-treino para a nossa língua [22]. Como esperado, o PTT5, por já ser um modelo pré-treinado para o português, teve desempenho bem superior.

O LegalSumm [25] também é um sumarizador aplicado a documentos em português, com foco em textos jurídicos utilizando *Tranformers* e implicação textual (*textual entailment*). A abordagem, proposta pelos autores, cria oito pedaços de texto a partir da decisão judicial e gera um sumário-candidato para cada. Em seguida, é avaliada a implicação textual entre a decisão e o sumário para seleção daquele candidato que obtiver a maior pontuação [25].

Este trabalho busca combinar os avanços recentes na área de sumarização extrativa, abstrativa e híbrida aplicados a documentos longos usando Modelos de Linguagem baseados em *Transformers* e seus desafios com o desafio adicional de lidar com documentos

longos em português jurídico.

Capítulo 3

Fundamentação Teórica

Este capítulo apresenta os aspectos técnicos mais importantes considerados neste trabalho para a sumarização abstrativa automática de texto com o uso de técnicas de PLN aplicadas a textos em português com conteúdo jurídico.

3.1 *Transformers*

Vaswani et al. [11] propuseram um modelo de arquitetura baseado inteiramente em um mecanismo de atenção e sem recorrência, em contraste com os modelos Seq2Seq, que usam as RNNs baseadas em LSTM [26] ou GRU [27]. Como o modelo não é recorrente, ele aumenta seu paralelismo, o que aumenta a eficiência computacional durante o treino [11].

Além da eficiência computacional, esse modelo de arquitetura estabeleceu o estado da arte em uma tarefa de geração de linguagem natural, *Neural Machine Translation* (NMT), mostrando ganhos de eficiência e qualidade sobre os modelos recorrentes anteriores baseados em Seq2Seq.

A arquitetura proposta por Vaswani et al. [11] é composta de um conjunto de *encoders* e *decoders* com *self-attention*, conforme Figura 3.1.

O *encoder*, terceiro bloco da esquerda para direita na Figura 3.1, é composto por uma pilha de N camadas idênticas. Cada camada tem duas subcamadas. A primeira é o mecanismo de *Multi-head Self-attention*, explicado abaixo, enquanto que a segunda é uma rede *feed-forward* simples e totalmente conectada. Além disso, há uma conexão residual [28] em torno de cada uma das duas subcamadas, seguida de normalização [11, 29].

O *decoder*, quarto bloco da esquerda para direita na Figura 3.1, é composto por uma pilha de N camadas idênticas. Além das duas subcamadas, conexões residuais e normalização, como o *encoder*, o *decoder* possui uma terceira subcamada, que é a *Multi-Head Cross Attention* na saída da pilha de *encoders*. A subcamada *Multi-head Self-attention* do

decoder é modificada para ter um mascaramento adicionado e, assim, evitar que as posições atendam às posições subsequentes e preservem a autorregressividade do modelo [11].

A entrada do *encoder* e do *decoder* é composta por *embeddings*. Estes são vetores que representam o texto, o qual foi segmentado em *tokens* e convertidos para uma representação vetorial numérica com a dimensão de tamanho d_{model}. Os *embeddings* são aprendidos pelo modelo para melhor representar os *tokens*, de forma a aproximar-se da semântica das palavras [11].

Positional Encodings são utilizados para compensar a ausência de recorrência ou convolução no *Transformer* para que o modelo saiba identificar a ordem dos *tokens* na sequência do texto de entrada. Para isso, são utilizadas duas funções senoidais para fazer esse *encoding* posicional, conforme Equações 3.1 a 3.2:

$$PE(pos, 2i) = sin(pos/10000^{2i/d_{model}}) \tag{3.1}$$

$$PE(pos, 2i+1) = cos(pos/10000^{2i/d_{model}}) \tag{3.2}$$

Onde *pos* é a posição do *embedding* do *token* e i é a dimensão. Cada dimensão do *positional encoding* corresponde a uma senoide.

O atenção, segundo bloco da esquerda para direita na Figura 3.1, que está contida dentro do *encoder* e do *decoder*, é uma função que mapeia *queries* (vetores Q) a um conjunto de chaves (vetores K) e valores (vetores V). A saída é calculada como uma soma ponderada dos valores [11]. O *Scaled Dot-Product Attention* é calculado sobre os vetores Q, K e V de acordo com a Equação 3.3:

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{3.3}$$

A variável d_k é a dimensão dos vetores de entrada K e Q. K^T é o vetor K transposto para a compatibilizar as dimensões dos vetores para a operação.

3.1.1 Mecanismos de atenção

Um dos pilares do modelo de *transformer* foi seu mecanismo de atenção [11]. Esse mecanismo compensou a m cada camada, que são responsáveis por apontar quaisção da recorrência do modelo, pois passou a receber a entrada completa e vê-la como um todo, ao contrário dos modelos Seq2Seq, e se orientar pelas várias *attention heads* em cada camada, que são responsáveis por apontar quais *tokens* receberão mais foco do modelo para cada *token* da entrada. Clark et al. [31] detalham algumas das funções aprendidas pelas

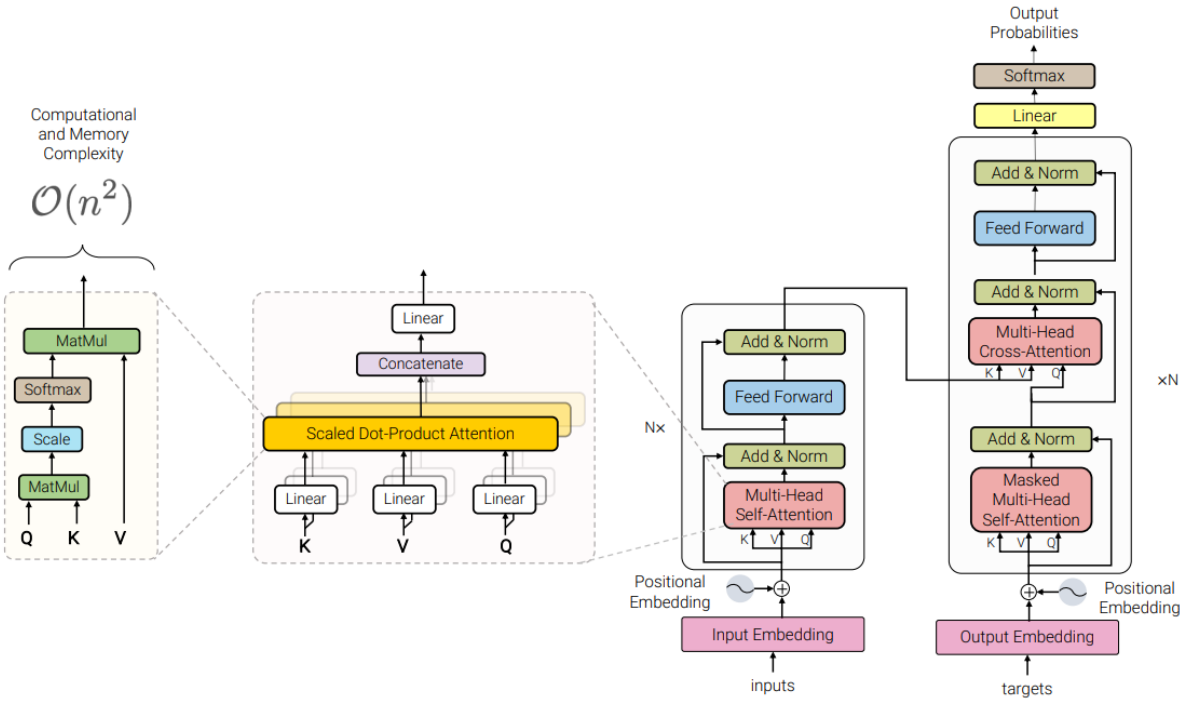

Figura 3.1: *Transformers* originais. Detalhamento do mecanismo de atenção (Fonte: [30]).

cabeças de atenção no modelo BERT, que são relacionadas a aspectos linguísticos, como: objetos diretos, artigos, adjetivos, preposições, pronomes, entre outros.

Esse mecanismo, conforme implementado por Vaswani et al. [11] limita o tamanho de textos que podem ser consumidos pelo modelo, pois a complexidade do mecanismo de atenção cresce quadraticamente com o tamanho da entrada, o que faz com que isso seja um fator limitante para a sumarização de textos longos.

A matriz de atenção, conforme ilustrado na Figura 3.2, para a entrada de exemplo *"The firm for which Jacob worked sent him to New York"*, gera uma matriz N x N onde cada *token* de entrada tem um peso de importância para todos os demais *tokens* de entrada, incluindo ele mesmo.

O *Scaled Dot Product Attention* da Figura 3.1 é calculado pela seguinte equação [11]:

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{3.4}$$

Especificamente, a parte QK^T da Equação 3.4 sozinha já consome $O(n^2)$ de tempo e memória [30], onde n é o tamanho da entrada.

Figura 3.2: Matriz de atenção quadrática do *Transformer* original (Fonte: [32]).

3.1.2 Mecanismos de atenção otimizados

Para escalar os modelos baseados em *transformers* para uso com documentos longos, foram propostos vários modelos ao longo dos anos endereçando o problema da atenção quadrática [30] com o intuito de melhorar sua relação com o tamanho da entrada. Os modelos aplicados à sumarização que obtiveram os melhores resultados em documentos longos usando atenção linear foram o *Longformer-Encoder-Decoder* (LED) [6] baseado no modelo do BART [8] e o *Big Bird* [10] baseado no modelo do *Pegasus* [9] que já chegaram a estabelecer o estado da arte para sumarização em alguns *datasets* de documentos longos como ArXiv [33], PubMed [33] e o BigPatent [34] quando foram desenvolvidos.

3.1.3 Modelo BART

O BART é um modelo baseado em *transformers* e foi pré-treinado seguindo os passos do RoBERTa [8, 21]. Este modelo tem como tarefa de pré-treinamento o *denoising*, que consiste em adicionar um ruído (*noising*) ao texto de entrada e depois reconstruí-lo corretamente [8]. Ele é treinado em duas tarefas que adicionam ruído à entrada, o *Text Infilling* e o *Sentence Permutation*, conforme pode ser ilustrado na Figura 3.3.

O *Text Infilling* mascara uma quantidade de trechos de texto variável, obtidos por uma Distribuição de Poisson ($\lambda = 3$), enquanto que o *Sentence Permutation* embaralha a ordem das sentenças para que o modelo reconstrua a entrada original antes do corrompimento pelas duas funções de ruído.

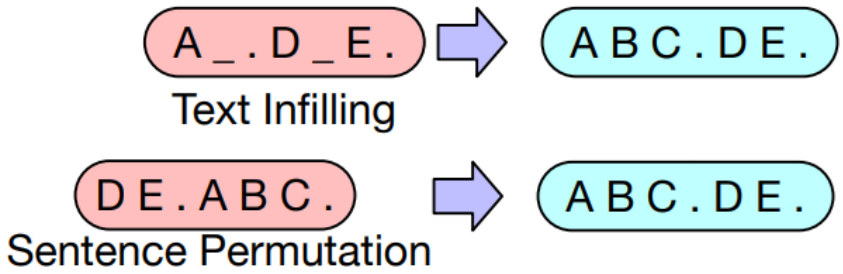

Figura 3.3: Tarefa de *denoising* para pré-treino do BART (Fonte: [8]).

Ele possui duas versões, a *base* com seis camadas de *encoder* e seis camadas de *decoder*. A versão *large* tem doze camadas de *encoder* e doze de *decoder* [8]. A versão *large* tem tido bons resultados para resumos abstrativos [2] e é a base para construção do modelo LED [6].

3.1.4 Modelo *Longformer-Encoder-Decoder* (LED)

O LED é uma extensão do modelo BART em que há uma modificação apenas das cabeças de atenção para que o modelo não cresça de maneira quadrática com o aumento da entrada [6]. O BART consegue ingerir até 1024 *tokens* e para chegar aos 16.384 *tokens*, o LED se utiliza deste artifício de linearização [6] para poder ingerir documentos longos sem que isso resulte em aumento significativo do tamanho do modelo.

Beltagy et al.[6] propõem uma abordagem para mitigar o problema da atenção quadrática, conforme ilustrado na Figura 3.4. Essa abordagem consiste no uso de uma janela deslizante de atenção local combinada com a atenção global. A parte em branco da figura representa o que não é mais necessário para o modelo, resultando em menor uso de *hardware* para treinamento e predição, sem que haja perda de qualidade no resultado final [6].

3.2 Modelos de Linguagem

Com o surgimento dos *transformers*, vários modelos baseados neles foram surgindo ao longo dos anos e estabelecendo novo estado da arte em várias tarefas de PLN como o BERT [20], que pré-treinou um modelo base nas tarefas de *Masked Language Modeling* (MLM) e *Next Sentence Prediction* (NSP) para, em seguida, fazer o ajuste fino em tarefas específicas sem a necessidade de retreiná-lo do zero [20]. Ele foi um dos que pavimentou o caminho de uso de *transfer learning* a partir de *transformers*.

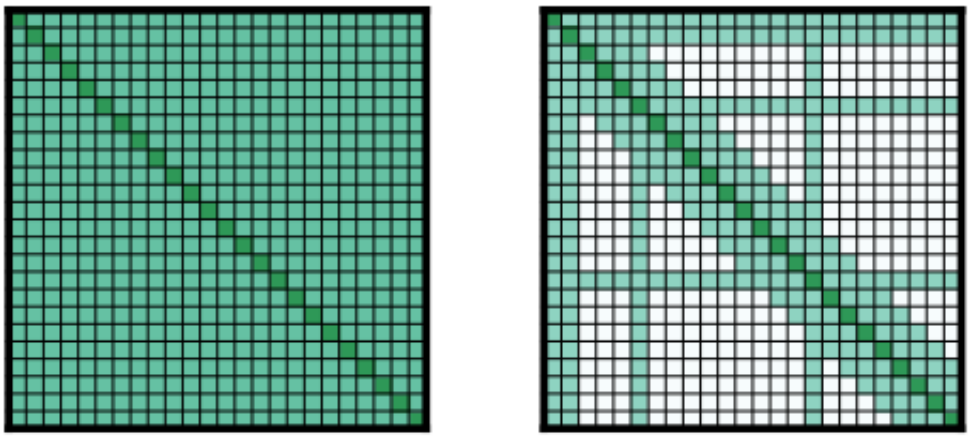

Figura 3.4: Mecanismo de Atenção: comparativo entre *transformer* original e LED (Fonte: [6]).

Os modelos BERT [20], BART [8], Pegasus [9], ROBERTa [21], T5 [23], entre vários outros, são exemplos de modelos de linguagem baseados em *transformers* que surgiram nos últimos anos. Novos modelos de linguagem que estão surgindo mais recentemente aumentaram bastante a quantidade de parâmetros e estão send classificados como Modelos de Linguagem Grandes, como o GPT-3 [35], o ChatGPT 4k [15] e o ChatGPT 16k [36], para mencionar apenas alguns dos mais proeminentes. Esses modelos de linguagem de grande escala estão alcançando novos patamares de dimensão, com quantidades de parâmetros variando de bilhões a trilhões.

3.3 MemSum

O MemSum é um modelo de sumarização extrativa que usa aprendizado por reforço ao invés de aprendizado supervisionado. Este modelo utiliza a métrica Rouge como função de recompensa para o aprendizado [16].

Para que o modelo possa ter a recompensa calculada de forma a que ele aprenda a maximizar a Rouge, é feito um pré-processamento no *dataset* que será usado para treino. O *dataset* é quebrado em sentenças e é feito uma comparação de todas as combinações de sentenças de entrada com as sentenças de saída (resumos), de forma a escolher a melhor combinação (usando *Beam Search*) das sentenças da entrada que maximizam o Rouge, de acordo com o resumo esperado. Com isso, o modelo tem o máximo de Rouge que ele pode atingir para cada exemplo, calculado usando *Beam Search* [16].

O modelo utiliza *GloVe Embeddings* [37] para representar o texto e tem três grandes camadas: *Local Sentence Encoder* (LSE), *Global Context Encoder* (GCE), *Extraction History Encoder* (EHE) e o *Extractor*, conforme mostrado na Figura 3.5.

No LSE, para cada sentença de entrada, as palavras são convertidas em *GloVe Embeddings* que são passados, internamente, para uma rede LSTM bidirecional. Essa rede, em conjunto com o *Multi-Head Pooling* (MHP), mapeia a entrada para *embeddings* de sentença [16]. O MHP é responsável por condensar a representação dos *embeddings* da sentença em um tamanho fixo, independente do tamanho original dela [38].

O GCE recebe os *embeddings* das sentenças do LSE e passa por uma rede LSTM bidirecional para produzir para cada sentença um *embedding* que codifica informações contextuais globais como como a posição da sentença no texto e informações sobre sentenças vizinhas [16].

O EHE faz o *encoding* das informações de histórico de extração e produz seus *embeddings* hs_i^r para cada sentença remanescente s_i^r. O EHE é composto por Nh camadas idênticas. Cada camada possui duas *Multi-Head self-Attention* (MHA). A MHA_1 é aplicada aos *embeddings* de sentenças remanescentes enquanto a MHA_2 é aplicada aos *embeddings* de sentenças extraídas. A saída das duas subcamadas de atenção captura a informação contextual de ambos os vetores: sentenças extraídas e remanescentes. A saída do EHE representa os *embeddings* do histórico de extração de sentenças para cada sentença extraída [16].

O *Extractor* é a última camada do agente e recebe as i saídas concatenadas para cada uma das sentenças remanescentes (r) do LSE (ls_i^r), do GCE (gs_i^r) e do EHE (hs_i^r). Esses insumos são utilizados para calcular a pontuação de cada sentença remanscente (us_i^r) e o sinal de parada de extração (P_{stop}).

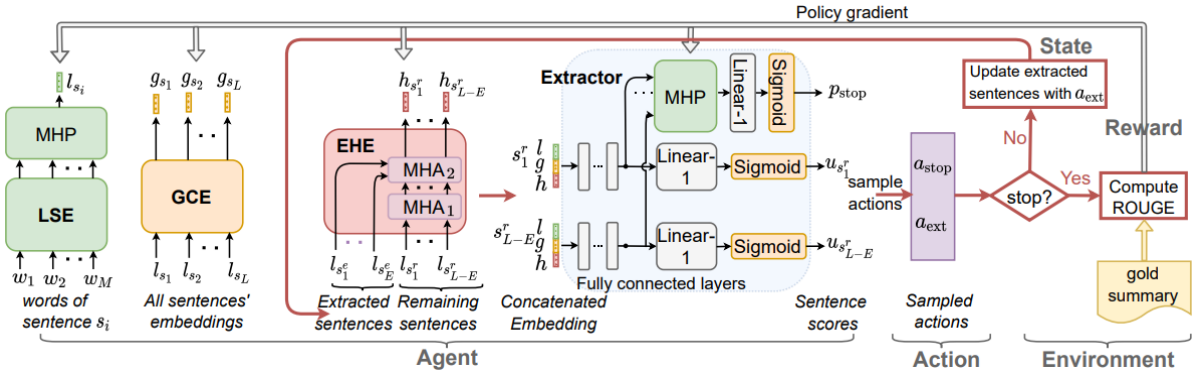

Figura 3.5: Arquitetura do Modelo MemSum (Fonte: [16]).

3.4 ChatGPT

O ChatGPT, ou GPT-3.5 Turbo, é um Modelo de Linguagem Grande que foi lançado pela OpenAI em novembro de 2022 [15]. Este modelo é o sucessor do GPT-3 [35] e foi

pré-treinado em um grande volume de textos para posterior ajuste fino em conversação usando *Reinforcement Learning from Human Feedback* (RLHF) [15]. Atualmente ele está disponível em duas versões, uma com 4.096 *tokens* de limite para a soma de entrada e saída do modelo e outra com limitação de 16.384 *tokens* para a soma de entrada e saída. A versão de 16.384 *tokens* foi lançado há poucos dias [36] e ainda não está disponível na nuvem da Microsoft, a Azure, que foi usada neste trabalho para integrar com o ChatGPT.

3.5 Métricas para avaliação do modelo proposto

A avaliação de textos gerados por máquinas é um problema aberto e objeto de pesquisa com uma variedade de métricas, e recompensas para algoritmos de aprendizado por reforço, sendo propostas ao longo dos anos na tentativa de automatizar as avaliações e depender cada vez menos da cara avaliação humana [2, 39].

Abaixo serão detalhadas as principais abordagens para avaliar sumarização abstrativa de textos, que envolvem desde a avaliação manual (humana), passando por avaliações automatizadas do sumário até a avaliação por feedback (recompensa), que se propõe a equilibrar os custos e benefícios dos dois extremos.

3.5.1 Avaliação Humana

A avaliação humana, apesar de conter subjetividade e possível inconsistência entre os critérios usados pelos avaliadores, é a que traz os melhores resultados, pois consegue avaliar critérios que são de difícil automatização como: fluência e legibilidade (ligados à qualidade do texto e acurácia), adequação, relevância e corretude factual (ligados à pertinência do sumário gerado em comparação com o texto original) [39]. As métricas automatizadas propostas na literatura procuram demonstrar sua efetividade através de alta correlação com a avaliação humana [12, 13, 14]. Isso demonstra seu efetivo valor e importância como *benchmark*.

Apesar de seu alto valor, essa abordagem traz consigo algumas dificuldades, como ganho de escala, pois a avaliação é feita manualmente e não é generalizável. Além disso, há dificuldades relacionadas à alta variância das avaliações por diferentes pessoas [39].

3.5.2 Rouge

Recall-Oriented Understudy for Gisting Evaluation (Rouge) [12] é a métrica mais utilizadas para avaliação na sumarização abstrativa de texto [2, 17] como também a mais antiga entre as elencadas abaixo. A abordagem usada pelo Rouge é de calcular a co-ocorrência de palavras entre o resumo de referência e o resumo gerado pelo algoritmo. A co-ocorrência é

calculada para *precision*, *recall* e *F1-score* e este último é o que é utilizado para avaliação dos resultados.

A co-ocorrência aparece na sumarização em três variações [2, 17, 39]: pode ser por n-grama, como usado em (i) Rouge-1: co-ocorrência de unigramas entre os resumos e (ii) Rouge-2: co-ocorrência de bigramas entre os resumos ou usando *Longest Common Subsequence* (LCS)[1], como em (iii) Rouge-L (co-ocorrência da maior subsequência de palavras entre os resumos). As comparações para sumarização, aplicam algoritmo de *stemming* como um pré-processamento.

Apesar de ser amplamente utilizada, esta métrica possui algumas desvantagens para a sumarização, uma vez que se baseia em comparações léxicas em vez de semânticas. Isso implica que, se as palavras aparecerem em uma ordem diferente no resumo, sem alterar seu significado, a métrica irá penalizar como um erro, o que é um equívoco. Além dessa possibilidade, por não considerar a semântica das palavras, sinônimos também são considerados erros, mesmo que o sentido do resumo seja inalterado. Uma forma de demonstrar este problema é através do uso de paráfrases, pois apesar do significado ser o mesmo que o texto original, esta métrica dará um *score* baixo em decorrência da diferença léxica.

Para usar esta métrica em textos em português, é necessário trocar o algoritmo de *stemming* usado e que é apropriado para o inglês para outro mais apropriado para o português.

3.5.3 MoverScore

O MoverScore [13], ao contrário da Rouge, não faz uma comparação léxica entre os textos. Esta métrica usa uma representação dos textos baseada em *embeddings* e os compara quanto à sua similaridade para mensurar o quanto um desvia do outro [13].

O MoverScore foi aplicado pelo autor em várias tarefas de geração de linguagem natural, entre elas, ele foi aplicado aos seguintes *datasets* de sumarização com resumos bastante curtos que possuem menos que 100 palavras[2]: TAC-2008 e TAC-2009.

Os melhores resultados obtidos pelas variações da métrica, descritas acima, quando aplicados à sumarização desses *datasets* foram: WMD-1 (*Word Mover's Distance* - função de distância entre textos de documentos usando unigrama) + BERT (mecanismo de *embedding*) + MNLI[3] (ajuste fino utilizado no BERT) + PMEANS (técnica de agregação utilizando as últimas cinco camadas do BERT e usando *power means*, que é uma média dos *embeddings* ponderada por uma exponenciação para agregar as *layers* do BERT) [13].

[1] LCS é a maior subsequência comum, não necessariamente contínua, entre dois textos

[2] `https://tac.nist.gov/`

[3] `https://cims.nyu.edu/~sbowman/multinli/`

Apesar da sua expressividade, esta métrica varia a melhor configuração de acordo com a tarefa de geração de linguagem natural e para cada *dataset* utilizado [13], o que torna seu uso um pouco mais complexo, pois para aplicá-la, será necessário ou escolher uma como padrão ou testar suas variações para encontrar aquela que possui a maior correlação com a avaliação humana.

Esta métrica está ganhando popularidade atualmente para avaliação de sumarização abstrativa de texto [2].

Para usar esta métrica em textos em português, é necessário substituir o BERT original, treinado para o inglês por outro treinado para o português, como o BERTimbau [40] e realizar o ajuste fino no MNLI original traduzido para o português[4].

3.5.4 BERTScore

O BERTScore [14], ao contrário do Rouge, não faz uma comparação léxica entre os textos. Esta métrica, assim como o MoverScore, usa uma representação contextualizada dos textos, mas difere na maneira como executa a comparação entre eles. O BERTScore usa uma abordagem gulosa para fazer o alinhamento entre os *tokens* dos textos e compara aqueles que possuem a menor distância de cosseno. Cada *token* pode ter sua importância ponderada pelo *Inverse Document Frequency* (IDF) do *token* no *corpus* ou não, dependendo da variação escolhida da métrica. O BERTScore, assim como o Rouge, usa *precision*, *recall* e *F1-score* e este último é o que é utilizado para avaliação dos resultados. Uma das vantagens do BERTScore é que ele é agnóstico sobre qual tarefa ele será aplicado e, por consequência, mais fácil de usar [14].

O BERTScore foi testado pelo autor com variações além do uso ou não do IDF. Ele obteve os melhores resultados em língua inglesa usando o $\text{RoBERTa}_{\text{LARGE}}$ [21], para o Chinês, foi o $\text{BERT}_{\text{BASE}}$ pré-treinado em Chinês obteve o melhor resultado, mas para as demais línguas, foi recomendado apenas o uso do $\text{BERT}_{\text{BASE}}$ multilingual [14].

Para usar esta métrica em textos em português, é necessário substituir o BERT original, treinado para o inglês por outro treinado para o português, como o BERTimbau [40] ou o multilingual sugerido pelo autor.

Esta métrica também está ganhando popularidade atualmente para avaliação de sumarização abstrativa de texto [2].

[4] `https://huggingface.co/datasets/dlb/plue`

Capítulo 4

Solução Proposta

Este capítulo detalha a solução proposta de solução desde o treinamento de modelos de sumarização até a forma como será integrada ao processo de trabalho do auditor do TCU. O foco da solução é que haja ganho de tempo para o auditor através da leitura dos resumos de texto gerados automaticamente pelo modelo de aprendizado de máquina ao invés da leitura de grande parte do documento, como é feito usualmente.

4.1 Experimentos Preliminares

Os objetivos destes experimentos iniciais foram os seguintes:

- Verificar se os modelos disponibilizados pelos autores já trariam bons resultados sem a necessidade de pré-treino ou ajuste fino;
- Reduzir a quantidade de modelos, filtrando e deixando apenas que os que se mostrassem mais promissores e com melhor entendimento da nossa língua.

Foram realizados alguns experimentos iniciais para verificar como os quatro modelos (BART, *Pegasus*, LED e *Big Bird-Pegasus*) se comportariam ao tentar sumarizar a seção de EXAME TÉCNICO de uma instrução pequena do TCU com conteúdo em português jurídico (Anexo I), sem nenhum pré-treino de Modelo de Linguagem em português ou mesmo ajuste fino em *datasets* de sumarização em português.

Para os experimentos, foi utilizada a seção de EXAME TÉCNICO da Instrução Processual do Anexo I, que gerou uma entrada de 3.728 *tokens* para os modelos. Esta entrada ultrapassa o limite suportado tanto pelo BART [8] quanto pelo *Pegasus* [9], que é de 1.024 *tokens*. Para que eles pudessem ser testados, a entrada foi truncada no limite máximo deles. O LED [6] e o *Big Bird-Pegasus* [10], que suportam 16.384 e 4.096 *tokens* respectivamente, puderam processar o texto por completo sem problemas.

As subseções abaixo mostram as diferentes estratégias utilizadas para os experimentos preliminares para seleção dos modelos iniciais para sumarização: (i) Texto original na entrada sem pré-processamento para sumarização, (ii) Texto normalizado na entrada para sumarização e (iii) Texto traduzido do português para inglês, sumarizado e traduzido de volta.

4.1.1 Texto original na entrada sem pré-processamento para sumarização

Usando o texto original, os modelos se comportaram da seguinte forma:

- **Sumarização do BART**: Resumo muito curto. Deixou uma quantidade grande de informações importantes de fora. Não se consegue entender o contexto geral da peça com este resumo.

  Resumo:

  O reitor da universidade, inicia a sua argumentação (peça 60) informando que o laudo apresentado pela representante apresenta divergências de informações. Destaca que o castelo d'água deverá ser demolido e reconstruído, devido ao desaprumo.

- **Sumarização do *Pegasus***: Erro de representação de diacríticos em **negrito** e resumo muito curto. Deixou uma quantidade grande de informações importantes de fora. Não se consegue entender o contexto geral da peça com este resumo.

  Resumo:

  O reitor da universidade, inicia a sua **argumentao** (**pea** 60) informando que o laudo apresentado pela representante apresenta **divergncias** de **informaes**.

- **Sumarização do LED**: Erro de iniciais maiúsculas em início de frase em **negrito** e resumo com conteúdo repetitivo.

  Resumo:

  o reitor da universidade, inicia a sua argumentação (peça 60) informando que o laudo apresentado pela representante apresenta divergências de informações, se comparados os percentuais levantados por ela ao estado real da obra.

  informa que o laudo apresentado pela representante apresenta divergências de informações, se comparados os percentuais levantados por ela ao estado real da obra.

  informa que o laudo apresentado pela representante apresenta divergências de informações, se comparados os percentuais levantados por ela ao estado real da obra.

informa que encaminhou em anexo documentos que comprovam essa condição e os problemas ocorridos ao longo do contrato.

informa que o castelo d'água deverá ser demolido e reconstruído, devido ao desaprumo observado, perceptível a olho nu. outro item reportado é a instalação de elevador, que não atendeu às especificações técnicas de projeto e que não foi aceito pela última fiscalização.

o reitor também afirma que faltam equipamentos na subestação, como disjuntores de alta tensão, quadros, barramentos, gabinete e demais elementos necessários ao funcionamento do equipamento.

acrescenta que está em andamento o processo de apuração de responsabilidades pelos problemas ocorridos, sendo que o TCU está acompanhando esses procedimentos, por meio do TC 001.267/2019-1, de responsabilidade da Secex Educação.

- **Sumarização do *Big Bird-Pegasus***: Modelo alucinou e ignorou a entrada, gerando resumo sem conexão alguma com o texto de entrada e, além disso, em outro idioma, o inglês.

  Resumo:

  in this paper, we investigate the properties of a system consisting of two coupled oscillators, one of which acts as a heat bath, and the other one as a mechanical resonator.

  we consider the case when the temperature of the system is lower than the critical temperature of the bath.

  we show that, in this case, the system behaves as a brownian particle.

  we also show that the brownian particle does not propagate in the bath.

  the aim of this work is to investigate the properties of a system consisting of two coupled oscillators, one of which acts as a heat bath, and the other one as a mechanical resonator. in this case, we consider the case when the temperature of the system is lower than the critical temperature of the bath.

  we show that, in this case, the system behaves as a brownian particle.

  the aim of this work is to investigate the properties of a system consisting of two coupled oscillators, one of which acts as a heat bath, and the other as a mechanical resonator. in this case, we consider the case when

  the temperature of the system is lower than the critical temperature of the

4.1.2 Texto normalizado na entrada para sumarização

Usando o texto normalizado, isto é, com a remoção de diacríticos, os modelos se comportaram da seguinte forma:

- **Sumarização do BART**: O modelo gerou a saída sem os diacríticos, como recebeu na entrada, mesmo que a palavra não exista em português (em **negrito**). Além disso, o resumo gerado é muito curto.

  Resumo:

  O reitor da universidade, inicia a sua **argumentacao** (peca 60) informando que o laudo apresentado pela representante apresenta **divergencias** de **informacoes**. Destaca que o castelo **d'agua devera** ser demolido e **reconstruido**, devido ao desaprumo observado.

- **Sumarização do *Pegasus***: O modelo gera a saída sem os diacríticos, como recebeu na entrada, mesmo que a palavra não exista em português. Gerou também palavras incompletas ou erradas (levant e pareceress) e trechos incompletos (trecho: "apresentado representante" seria "apresentado pela representante") todos em **negrito**.

  Resumo:

  O reitor da universidade, inicia a sua **argumentacao** (**peca** 60), informando que o laudo **apresentado representante** apresenta **divergencias** de **informacoes**, se comparados os percentuais **levant** por ela ao estado real da obra.

  Tambem indicou os anexos com os referidos **pareceress**.

- **Sumarização do LED**: O modelo gera a saída sem os diacríticos, como recebeu na entrada, mesmo que a palavra não exista em português. Erro de iniciais maiúsculas em início de frase e palavras incompletas em **negrito**. Resumo bastante repetitivo.

  Resumo:

  o reitor da universidade, inicia a sua **argumentacao** (**peca** 60) informando que o laudo apresentado pela representante apresenta **divergencias** de **informacoes**, se comparados os percentuais levantados por ela ao estado real da obra.

  acrescenta que foi realizado levantamento **topografico** para a **confirmacao** do desaprumo e os resultados foram apresentados a engenheiro do departamento de engenharia, que informou haver necessidade de **demolicao**, pelo risco de tombamento do **reservatorio**, se for preenchido. Destaca que o castelo **d'agua** devera ser demolido e **reconstruido** , devido ao desaprumo observado , **perceptivel** a olho nu. **o** reitor da universidade, inicia a sua **argumentacao** (**peca** 60) informando que

o laudo apresentado pela representante apresenta **divergencias** de **informacoes**, se comparados os percentuais levantados por ela ao estado real da obra.

acrescenta que foi realizado levantamento **topografico** para a **confirmacao** do desaprumo e os resultados foram apresentados a engenheiro do departamento de engenharia , que informou haver necessidade de **demolicao** , pelo risco de tombamento do **reservatorio**, se for preenchido.

o reitor **tambem** afirma que faltam equipamentos na **subestacao**, como disjuntores de alta **tensao**, quadros, barramentos, gabinete e demais elementos **necessarios** ao funcionamento do equipamento. **o** reitor da universidade , inicia a sua **argumentacao** (**peca** 60) informando que o laudo apresentado pela representante apresenta **divergencias** de **informacoes**, se comparados os percentuais levantados por ela ao estado real da obra.

acrescenta que esta em andamento o processo de **apuracao** de responsabilidades **pel**

- **Sumarização do *Big Bird-Pegasus***: Modelo alucinou e ignorou a entrada, gerando resumo sem conexão alguma com o texto de entrada e, além disso, em outro idioma, o inglês.

  Resumo:

  in this paper, we address the question of whether or not the concept of mass is compatible with the notion of causality.

  in particular, we investigate whether or not the mass is compatible with the notion of causality.

  the study of causality is one of the most important topics in modern physics.

  the concept of causality was introduced by einstein more than half a century ago.

  nowadays, the concept of causality is widely used in various fields of science. in particular, the concept of mass

  was introduced in physics more than half a century ago. in this paper, we address the question of whether or not the concept of mass is compatible with the notion of causality.

  in particular, we investigate whether or not the mass is compatible with the notion of causality. the concept of mass was introduced by einstein more than half a century ago.

  the concept of mass was introduced in physics more than half a century ago.

  nowadays, the concept of mass is widely used in various fields of science.

the study of causality is one of the most important topics in modern physics.

the concept of mass was introduced in physics more than half a century ago.

4.1.3 Texto traduzido do português para inglês, sumarizado e traduzido de volta

Usando o texto traduzido para o inglês, sumarizado e traduzido de volta para o português, os modelos se comportaram da seguinte forma:

- **Sumarização do BART**: Erro de tradução (onde há "parte" seria "peça" e onde há "assume" seria "retoma") em **negrito**.

  Resumo:

  O reitor da universidade inicia seu argumento (**parte** 60) afirmando que o relatório apresentado pelo representante apresenta discrepâncias na informação.Na próxima **parte** (**parte** 61), o reitor informa o encaminhamento da planilha que contém as informações solicitadas pelo item 14.3.1 da ordem ministerial. Na peça 64, o reitor **assume** uma série de argumentos já enfrentados pelas instruções anteriores.

- **Sumarização do *Pegasus***: Erro de tradução ("Dean" foi confundido com nome próprio e não foi traduzido para "Reitor" em português) em **negrito**.

  Resumo:

  O relatório apresentado pelo representante apresenta discrepâncias nas informações, se as porcentagens levantadas por ela forem comparadas ao estado real do trabalho.

  O **Dean** também afirma que a subestação não possui equipamentos, como disjuntores de alta tensão, quadros de chave, barramentos, gabinete e outros elementos necessários para que o equipamento funcione.

  Os serviços relacionados à construção do Bloco 03 da IEAA ficaram paralisados em outubro de 2017, nesse período o bloco começou a ter problemas, naturais, de desgaste e depreciação devido à falta de manutenção, além de ter passado por eventos de roubo de roubo, depredação e infestação.

- **Sumarização do LED**: Erro de tradução (onde há "Nesta carta" seria "Neste relatório") e quebra de linha de parágrafo errada ("no trabalho.") em **negrito** e resumo repetitivo.

  Resumo:

  Nesta carta, o reitor da universidade inicia seu argumento afirmando que o relatório apresentado pelo representante apresenta discrepâncias nas informações, se

as porcentagens levantadas por ela forem comparadas ao estado real do trabalho. Além dos argumentos já desenvolvidos na carta anterior, destaca as diferenças de valores em relação à data base dos orçamentos (de 2014 a 2021).

Ele acrescenta que outro fator a ser considerado é o desconto oferecido pelo representante atual, de 13,68%, um desconto que não está incluído em um orçamento de referência baseado em sistemas oficiais.

Ele também destaca a existência de várias patologias no edifício, como infiltração, infestação por cupins e mofo, que causaram a perda de muitos serviços executados, além da ocorrência de roubo de elementos já concluídos **no**

trabalho.

Ele destaca que o castelo de água terá que ser demolido e reconstruído, devido ao desalinhamento observado, perceptível a olho nu.

O reitor da universidade inicia seu argumento afirmando que o relatório apresentado pelo representante apresenta discrepâncias em informações, se as porcentagens levantadas por ela forem comparadas ao estado real do trabalho.

Ele acrescenta que outro fator a ser considerado é o desconto oferecido pelo representante atual, de 13,68%, um desconto que não está incluído em um orçamento de referência baseado em sistemas oficiais.

Ele acrescenta que outro fator a ser considerado é o desconto oferecido pelo representante atual, de 13,68%, um desconto que não está incluído em um orçamento de referência baseado em sistemas oficiais.

Ele também destaca a existência de várias patologias no edifício, como infiltração, infestação por cupins e mofo, que causaram a perda de muitos serviços executados, além da ocorrência de roubo de elementos já concluídos **no**

trabalho.

Ele destaca que o castelo de água terá que ser demolido e reconstruído, devido ao desalinhamento observado, perceptível a olho nu.

Ele também destaca que a subestação não possui equipamentos, como disjuntores de alta tensão, placas de chave, barramentos, gabinete e outros elementos necessários para que o equipamento funcione.

Outro item relatado é a instalação de um elevador, que não atendeu às especificações do projeto técnico e não foi aceito pela última inspeção.

- **Sumarização do *Big Bird-Pegasus***: Modelo alucinou e ignorou em grande parte a entrada gerando texto sem sentido e com muita repetição, mesmo com a entrada traduzida para o inglês, língua em que ele foi treinado e obtém ótimos resultados [2].

  Resumo:

  Neste relatório, abordamos a questão de saber se o reitor da Universidade de Santiago de Compostela pode ser removido de sua posição e se o reitor pode ou não ser removido de sua posição.

  Também abordamos a questão de saber se o reitor da Universidade de Santiago de Compostela pode ser removido de sua posição e se o reitor pode ou não ser removido de sua posição.

  [[Seção]] O objetivo deste relatório é abordar a questão de saber se o reitor da Universidade de Santiago de Compostela pode ser removido de sua posição e se o reitor pode ou não ser removido de sua posição.

  [[Seção]] O objetivo deste relatório é abordar a questão de saber se o reitor da Universidade de Santiago de Compostela pode ser removido de sua posição e se o reitor pode ou não ser removido de sua posição.

  [[seção]] O objetivo deste relatório é abordar a questão de saber se o reitor da Universidade de Santiago de Compostela pode ser removido de sua posição e se

Nenhum dos modelos apresentou resultado satisfatório utilizando as versões disponibilizados pelos seus respectivos autores. Com isso, conclui-se que o treinamento de modelos é necessário para gerar resumos satisfatórios para o TCU. Além disso, entre os modelos que suportam documentos longos, LED e *Big Bird-Pegasus*, o primeiro teve resultados menos problemáticos e se mostrou mais promissor. Para os os próximos experimentos, os demais modelos foram descartados.

4.2 *Datasets* utilizados para experimentos

A sumarização de documentos é um problema complexo e tem seu resultado fortemente influenciado pelo *dataset* ao qual o modelo é exposto para treinamento, como pode ser observado na Seção 4.3. Ressalta-se que o TCU, no início deste trabalho não possuía nenhum *dataset* para treinamento de modelo de linguagem nem para sumarização e, por isso, foi necessário testar alguns já existentes para verificar como seria a sua transferência de conhecimento para os documentos do Tribunal. Abaixo são listados os *datasets* usados nos experimentos com suas características no que diz respeito à sumarização.

4.2.1 BRWac2Wiki - *Brazilian Portuguese Wikipedia Dataset for Summarization*

O BRWac2Wiki é um *dataset* de sumarização multi-documentos para documentos longos [22] inspirado no WikiSum [41] e possui cerca de 114 mil exemplos. Cada exemplo possui um título da Wikipedia, seu resumo e uma lista de documentos relacionados ao título. Esses documentos são textos de sites do corpus BrWac [42].

Cada exemplo de entrada traz o título do texto da Wikipedia separado por </s> e até 15 documentos (truncados até 100 palavras) que explicam o título e são separados por <\s>. Segue abaixo um exemplo de entrada, ilustrando seu formato [22]:

```
título do assunto do exemplo </s>
texto1 sobre o assunto <\s>
texto2 sobre o assunto <\s>
...
texto15 sobre o assunto <\s>
```

O resumo utilizado para predição é o primeiro parágrafo da Wikipedia sobre o assunto do título [22].

4.2.2 RulingBR - *Brazilian Legal Ruling Dataset for Summarization*

O RulingBR é um *dataset* de decisões judiciais do STF com documentos longos. Cada exemplo possui uma Ementa (resumo), um Relatório (uma compilação dos principais argumentos e eventos ocorridos durante o julgamento) e pelo menos um Voto. Os Votos devem abordar todos os pontos levantados pelos peticionários. É um *dataset* pequeno, com cerca de 10 mil exemplos antes de ser dividido em *datasets* de treinamento (cerca de 6 mil exemplos), validação (cerca de 2 mil exemplos) e teste (cerca de 2 mil exemplos) [43]. Este domínio está mais relacionado ao do TCU do que a Wikipedia e se mostra mais adequado para transferir o modelo treinado para resumir os documentos do TCU. Para cada exemplo, tem-se como entradas os Relatórios e Votos, ambos documentos longos, e como saída a Ementa do texto. Em geral, o Relatório representa cerca de 22% do conteúdo completo. [43] O Voto, poderá conter somente o voto do relator, caso os demais ministros concordem com o relator, ou votos individuais para cada ministro, caso contrário. Como os votos precisam abordar todos os pontos levantados pelos peticionários, essa tende a ser a maior seção, cobrindo cerca de 69% do conteúdo completo. [43] Ambos os documentos necessitam de truncamento, em geral, para ficar dentro do limite de entrada do modelo usado. A estratégia utilizada será abordada em cada caso.

4.2.3 BrWac - *Large Web corpus for Brazilian Portuguese*

O *dataset* BrWac é formado por mais de 60 milhões de páginas coletadas e filtradas, resultando em 3,53 milhões de documentos e 2,68 bilhões de tokens distribuídos em 145 milhões de sentenças [42]. Esse conjunto de dados torna possível o pré-treinamento de modelos de linguagem para compreender textos em português brasileiro, assim como foi feito com modelos como BERT [20] e T5 [23] para pré-treinar as versões em português: BERTimbau [40] e PTT5 [24], respectivamente.

4.2.4 TCU-LM - *TCU Legal Documents for Language Models Dataset*

O *dataset* TCU-LM é composto por pouco mais de 92 mil peças processuais do TCU, resultando em cerca de 6,2 milhões de sentenças e 223 milhões de *tokens*. Esse *dataset* torna possível o pré-treinamento de modelos de linguagem para compreender textos jurídicos em português brasileiro. Este *dataset* foi construído como parte desta pesquisa com o objetivo de que os modelos de sumarização pudessem entender o português jurídico, que é o que está presente nas peças processuais do TCU.

4.2.5 TCU-Summ - *TCU Legal Documents for Summarization Dataset*

Para testar os modelos com dados específicos do TCU, no âmbito desta pesquisa, foi gerado um pequeno *dataset* a partir das instruções processuais do TCU, removendo os primeiros parágrafos do texto, que correspondiam nos documentos selecionados a um resumo bem curto, para usá-los como saída, enquanto que o restante do texto era usado como entrada para o modelo ser treinado. O TCU-Summ possui cerca de 21,5 mil exemplos antes de ser dividido em *datasets* de treinamento (cerca de 17,5 mil exemplos), validação (cerca de 2 mil exemplos) e teste (cerca de 2 mil exemplos). Embora seja maior que o RulingBR, o texto é menos diversificado, com poucas variações de texto em comparação com o primeiro. Os resumos deste *dataset* são muito limitados em tamanho, com no máximo dois pequenos parágrafos.

4.3 Abordagens para construção da modelo de geração de resumo

Nesta seção, serão detalhadas as abordagens utilizadas para sumarização de textos longos com conteúdo jurídico no TCU. Cada subseção abordará as motivações do experimento,

os resultados e eventuais problemas que se manifestam no resumo gerado. Com relação à forma de geração de resumo, neste trabalho, foram experimentadas as gerações abstrativas, extrativas e híbridas.

Os experimentos que utilizaram a forma de geração abstrativa resumem apenas a seção EXAME TÉCNICO da Instrução Processual, com o objetivo de dar um foco maior na parte mais relevante do texto e evitar que partes menos importantes dele aparecessem no resumo. A seção de EXAME TÉCNICO é onde são descritas com detalhes as questões abordadas pela instrução, as normas legais envolvidas e as razões pelas quais cada proposição de encaminhamento foi determinada. Além da importância de seu conteúdo para o resumo, esta seção, em geral, corresponde a grande parte do documento e foi apontada por vários auditores, no questionário (Apêndice A), como uma seção relevante para resumo. Outra questão que motivou também o resumo apenas desta seção é a limitação de tamanho de texto para ingestão pelo modelo abstrativo, o que tornava necessário truncar o texto de entrada para documentos muito longos e, com isso, partes importantes do texto poderiam ser descartadas.

A Figura Figura 4.1 ilustra as estratégias experimentadas em ordem cronológica, destacando as abordagens, *datasets* e formas de geração de resumo utilizadas em cada etapa, as quais serão detalhadas nas subseções a seguir.

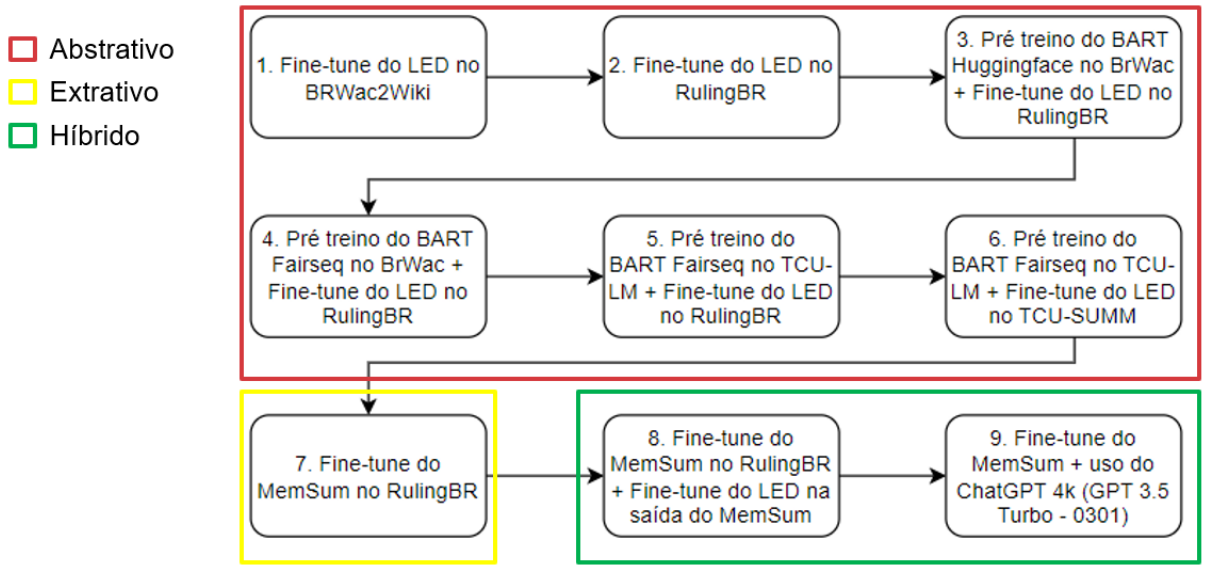

Figura 4.1: Abordagens experimentadas para geração de resumos.

4.3.1 Ajuste fino do LED no BRWac2Wiki

A primeira abordagem usou o modelo LED, que suporta 16.384 *tokens* de entrada e, com isso, foi possível resumir as instruções processuais (documentos longos) do TCU. O *dataset* escolhido para esta iteração foi o BRWac2Wiki, que é um *dataset* para sumarização de documentos longos com grande quantidade de exemplos para treinamento e com conteúdo diversificado de assuntos tratados [22].

A primeira variação desta abordagem com o modelo LED usou o *dataset* como disponibilizado por Oliveira e Costa [22] e treinou por aproximadamente 1 semana, de forma distribuída, utilizando 4 GPUs de alto desempenho (NVidia RTX 3090) no limite de suas capacidades.

Esta variação, aplicada à seção de EXAME TÉCNICO da Instrução Processual, não trouxe bons resultados, pois dada a estrutura do *dataset*, conforme descrito na Subseção 4.2.1, em que a entrada explica o título, fez com que o modelo aprendesse a focar apenas em explicá-lo, ignorando o conteúdo. Isso resultou em resumos que procuravam explicar o que era um EXAME TÉCNICO, quando que este título não possui uma semântica e é usado como uma seção marcadora da Instrução Processual.

Resumo da Instrução do Anexo I - BRWac2Wiki - original:

"o anexo técnico é um documento de análise técnicas de processos de anexação de documentos , que permite aplicar ao público a informações sobre os processos que ocorrem em um determinado edifício ou em uma área de construção , com o objetivo de identificar os fatores que influenciam sobre o desenvolvimento de uma obra ou edificado . os anexos têm sido utilizados em diversas áreas de pesquisa , como engenharia civil , engenho civil , tecnologia civil e engenheiro civil , entre outros . os documentos também podem ser encontrados em forma de biblioteca , em formato de cartão , ou como documentos anexados . o primeiro documento que foi encontrado no brasil foi publicado pela primeira vez em 17 de janeiro de 2005 . o documento foi escrito pela empresa brasileira de engenheiros vistoriadores (empresas vistoria) e foi emitido pela prefeitura do município de são paulo , no dia 1 de julho de 2005 , com a finalidade de conclarar os resultados da licenciamento do projeto ."

O texto gerado foi sem sentido e com vários erros de escrita. Além desses aspectos, o resumo ignorou o conteúdo da instrução, que era relacionado a uma obra em uma Universidade que havia sido paralisada. O resumo foi baseado quase que exclusivamente no título da seção, EXAME TÉCNICO. Para contornar esse problema de geração de texto baseado quase que exclusivamente no título, foi gerada uma segunda versão de *dataset* removendo o título de todos os exemplos para que o modelo fosse forçado durante o treinamento a se

basear apenas no conteúdo dos vários documentos que compõem cada exemplo, conforme detalhado na Subseção 4.2.1.

Esta variação teve um resultado ainda pior, gerando textos completamente sem sentido e, agora, sem considerar sequer o título para a geração do resumo. O título EXAME TÉCNICO também foi removido da instrução do Anexo I para que ficasse no mesmo formato que o LED foi treinado. O resumo, completamente desconectado do conteúdo do texto, pode ser visto abaixo:

Resumo da Instrução do Anexo I - BRWac2Wiki - sem título:

"ngc 001 é uma galáxia elíptica (e s0) localizada na direcção da constelação de eridanus . possui uma declinação atlântica de 11 09 09 e uma ascensão recta de 0 horas , 06 minutos e 07 , 8 segundos . a galáxia ngc 0001 foi descoberta em 1886 por frank leavenworth ."

Os modelos da variação 1 (*dataset* original com título) e variação 2 (*dataset* sem título) tiveram resultados muito bons no *dataset* BRWac2Wiki e seus resultados são superiores em algumas métricas aos publicados por Oliveira e Costa [22] no PLSUM, como pode ser visto na Tabela 4.1. A comparação serve apenas para mostrar que mesmo com uma performance boa no *dataset* de origem, ambos os modelos tiveram problemas para transferir o aprendizado do modelo para os textos jurídicos do TCU, como listados a seguir:

1. Geração de palavras inexistentes no português
2. Erros de concordância verbal e/ou nominal
3. Uso inadequado de iniciais maiúsculas e minúsculas
4. Alucinações extrínsecas presentes em todas as instruções testadas, isto é, o texto de entrada não dá suporte nem contradiz o que é gerado pelo resumo [3]

Tabela 4.1: Comparativo da Rouge para os modelos no BRWac2Wiki.

	Rouge-1	**Rouge-2**	**Rouge-L**
PLSUM	33.4	16.4	27.1
LED treinado COM título	33.9	14.0	30.7
LED treinado SEM título	34.8	14.5	31.5

4.3.2 Ajuste fino do LED no RulingBR

A segunda abordagem usou o mesmo modelo que a anterior, o LED. O *dataset* escolhido para esta iteração foi o RulingBR, que é um *dataset* para sumarização de documentos

longos com pouca quantidade de exemplos para treinamento e com conteúdo menos diversificado de assuntos tratados que o BRWac2Wiki. Apesar disso, este *dataset* tem conteúdo jurídico e é bem mais próximo do que se deseja resumir neste trabalho, que são as Instruções Processuais.

Com o objetivo de mitigar os problemas relacionados às alucinações que geraram resumos sem sentido nas instruções do TCU, além dos erros de escrita relacionados ao conteúdo jurídico, houve a troca para este *dataset*.

Foram experimentadas diversas variações deste modelo relativas à forma de truncamento da entrada (Votos e Relatórios). Estas incluíam: somente relatórios truncados até o limite de 16.384 *tokens*; somente votos truncados até o mesmo limite; e uma combinação dinâmica de relatórios e votos.

Esta última variação procurava dividir a entrada meio a meio quando ambos os componentes excediam 50% do tamanho total permitido. Caso um deles fosse menor, o outro era expandido para garantir que o modelo recebesse a máxima quantidade de informações para resumo.

Estas variações foram inspiradas pelo método utilizado no LegalSumm [25] para lidar com este mesmo *dataset*. A diferença reside apenas no fato de que o LED suporta uma quantidade de *tokens* de entrada muito maior que o LegalSumm.

Os resumos gerados em cada variação para a instrução do Anexo I são mostrados abaixo:

Resumo da Instrução do Anexo I - RulingBR - somente relatório:

"DIREITO ADMINISTRATIVO. AGRAVO REGIMENTAL EM RECURSO EXTRAORDINÁRIO. RESPONSABILIDADE OBJETIVA DO ESTADO. VERIFICAÇÃO DOS ELEMENTOS CONFIGURADORES. SÚMULA 279/STF. ANÁLISE DA OCORRÊNCIA DE EVENTUAL AFRONTA AOS PRECEITOS CONSTITUCIONAIS INVOCADOS NO APELO EXTREMO DEPENDE DA REELABORAÇÃO DA MOLDURA FÁTICA CONSTANTE DO ACÓRDÃO RECORRIDO. INOVAÇÃO RECURSAL: IMPOSSIBILIDADE. AC ARE 639.228-RG/SP, Rel. Min. Cezar Peluso, assentou que o referido dispositivo constitucional exige a explicitação, pelo órgão jurisdicional, das razões do seu convencimento, dispensando o exame detalhado de cada argumento suscitado pelas partes. A controvérsia, a teor do que já asseverado na decisão guerreada, não alcança estatura constitucional. Não há falar, nesse compasso, em afronta aos preceitos constitucionais indicados nas razões recursais, porquanto compreender de modo diverso exigiria análise da legislação infraconstitucional encampada na decisão prolatada pela Corte de origem, o que torna oblíqua e reflexa eventual ofensa, insuscetível, portanto, de viabilizar o conhecimento do recurso extraordinário. Desatendida a exigência do art. 102, III, "a", da

Lei Maior, nos termos da remansosa jurisprudência desta Corte. As razões do agravo regimental não se mostram aptas a infirmar os fundamentos que lastrearam a decisão agravada. Agravo regimental conhecido e não provido."

Resumo da Instrução do Anexo I - RulingBR - somente voto:

"EXAME TÉCNICO 7. O reitor da universidade, inicia a sua argumentação (peça 60) informando que o laudo apresentado pela representante apresenta divergências de informações, se comparados os percentuais levantados por ela ao estado real da também. Informa que encaminhou em anexo documentos que comprovam essa condição e os problemas ocorridos ao longo do contrato. Destaca que o castelo d'água deverá ser demolido e reconstruído, devido ao desaprumo observado, perceptível a olho nu. Acrescenta que foi realizado levantamento topográfico para a confirmação do desap 12o e os resultados foram apresentados a engenheiro do departamento de engenharia, que informou haver necessidade de demolição, pelo risco de tombamento do reservatório, se for preenchido. Também indicou os anexos com os referidos pareceres. 9. Outro item reportado é a instalação de elevador, que não atendeu às especificações técnicas de projeto e que não foi aceito pela última fiscalização. Por isso, os valores referentes à instalação foram incluídos entre as verbas que a construtora deverá restituir aos cofres da União. A peça contém um boletim de ocorrência autuado junto à polícia civil do Amazonas (p. 248-249). Precedentes. O TCU está acompanhando esses procedimentos, por meio do TC 001.267/2019-1, de responsabilidade da Secex Educação. Oforma que no processo de sindicância já evidenciado que os valores pagos à construtora não condizem com o que foi efetivamente executado. 12. Ressalta a existência de diversas patologias na edificação como infiltrações, infestação por cupins e mofo, que têm causado a perda de muitos serviços executados, além da ocorrência de furtos de elementos já concluídos da obra. Quanto a isso anexou à documentação o boletim de determinação do TCU no sentido de anular o RDC 1/2021. 15. Nessa resposta, o reitor colaciona texto redigido pela Diretora do Instituto de Educação, Agricultura e Ambiente – IEAA, Professora Ana Cláudia Fernandes Nogueira, uma vez que aquele instituto seria o mais impactado pela demora na conclusão das obras. Em resumo, a professora narra toda a história do instituto e descreve as dificuldades de atuação em edifício doado pela Prefeitura de Humaitá, que submete professores, alunos e funcionários a situações improvisadas, prejudicando o ensino e a pesquisa de maneira dramática, frisando que a não entrega do bloco 3, um objeto deste processo, continuará afetando os trabalhos administrativos e acadêmicos, bem como a qualidade do ensino prestado. 16. Por fim, esse reitor acrescenta também o depoimento do Diretor do Departamento HC condições atuais da edificação e a atualização dos referenciais de preços, pode-se afirmar que"

Resumo da Instrução do Anexo I - RulingBR - relatório combinado com voto:

"DIREITO ADMINISTRATIVO. SERVIDOR PÚBLICO. PLANILHA ORÇAMENTÁRIA. DESAPROMO DO RESERVATÓRIO D'ÁGUA. DEMORA NA CONCLUSÃO DAS OBRAS. AUSÊNCIA DOS REQUISITOS DO PERICULUM IN MORA E DO FUMUS BONI IURIS, QUE INCLUEM A PARALISAÇÃO CAUTELAR DO RDC 1/2021. 1. O desaprumo do reservatório d'água diz respeito à demora na conclusão das obras e ausência dos requisitos do periculum in mora e do fumus boni iuris, que indiquem a paralisação cautelar do RDC. 2. Considerando os serviços que efetivamente precisam ser executados, dadas as dificuldades de atuação em edifício doado pela Prefeitura de Humaitá, que submete professores, alunos e funcionários a situações improvisadas, prejudicando o ensino e a pesquisa de maneira dramática, frisando que a não entrega do bloco 3, obra objeto deste processo, continuará afetando os trabalhos administrativos e acadêmicos, bem como a qualidade do ensino prestado. 3. O reitor também afirma que faltam equipamentos na subestação, como disjuntores de alta tensão, quadros, barramentos, gabinete e demais elementos indispõe ao funcionamento do equipamento. 4. Os problemas com essa edificação já são antigos e certamente contou com a contribuição de muitos agentes. Os valores pagos à construtora não correspondem com o que foi executado. 5. Por fim, a peça contém um bolet"

Como pode ser observado na Tabela 4.2, as três variações foram bastante promissoras e a abordagem mais robusta ao transferir para as instruções foi a combinação dinâmica de relatório e voto quando aplicada a outros documentos além da instrução de exemplo.

Tabela 4.2: Comparativo da Rouge para os modelos no RulingBR.

	Rouge-1	**Rouge-2**	**Rouge-L**
LegalSumm	43	27	35
LED treinado (somente relatório)	50.6	33.2	48.8
LED treinado (somente voto)	48.9	30.5	47.0
LED treinado (relatório + voto)	51.1	30.7	49.0

A comparação serve apenas para mostrar que mesmo com uma performance boa no *dataset* de origem, todas as variações dos modelos tiveram algum tipo de problema para transferir o aprendizado do modelo para os textos jurídicos do TCU, como listados a seguir:

1. Geração de palavras inexistentes no português;
2. Foco em termos mais frequentes do *dataset*, mesmo que não existam na instrução, como "recurso extraordinário" e "agravo", além dos nomes de alguns ministros do STF apareceram em alguns casos;

3. Pequenas variações de texto como quebras de linha (\n) ou tabulações (\t) faziam com que o resumo fosse gerado bastante diferente;

4. Alucinações extrínsecas presentes em algumas instruções testadas, isto é, o texto de entrada não dá suporte nem contradiz o que é gerado pelo resumo [3].

Apesar das alucinações continuarem presentes nos documentos, foi observado nos experimentos que o conteúdo estava bem mais próximo do texto original, ao contrário do que ocorria com o *dataset* anterior, em que o documento inteiro era sem sentido e não relacionado ao texto de entrada. Esse *dataset* se mostrou mais apropriado e foi utilizado nos experimentos seguintes. O BRWac2Wiki foi descartado para os próximos experimentos.

4.3.3 Pré treino do BART Huggingface no BrWac com ajuste fino do LED no RulingBR

A terceira abordagem usou o mesmo modelo que a anterior, o LED e mesmo *dataset*, o RulingBR. Nesta abordagem, com a expectativa de mitigar os problemas relacionados às alucinações que geraram resumos com algumas partes não relacionadas ao documento de entrada, além dos erros de escrita relacionados ao conteúdo jurídico, foi adicionado um passo de pré-treino para que o modelo entendesse o português brasileiro.

O pré-treino foi realizado no início deste trabalho e nenhum dos modelos presentes na solução final (MemSum e ChatGPT) ainda existia, portanto, esta alternativa foi considerada a mais promissora para atender as necessidades deste trabalho.

Como detalhado na Seção 3.1.4, o LED é uma extensão do modelo BART pré-treinado com seus pesos originais, apenas com modificações nas cabeças de atenção do *transformer*. O pré-treino, de fato, ocorreu no BART, que no trabalho de Lewis et al. [8] seguiu os passos de pré-treino do RoBERTa [21].

Para realizar o pré-treino do BART como Modelo de Linguagem, este trabalho se baseou em trabalhos anteriores que fizeram a adaptação de modelos anteriormente pré-treinados para o inglês e que foram pré-treinados para o português, como o BERTimbau [40]. Os hiperparâmetros e demais detalhes específicos do modelo foram baseados no trabalho de Lewis et al. [8].

A biblioteca utilizada, Huggingface[1], não tinha nenhum suporte para pré-treino de Modelo de Linguagem para o BART nem para sua tarefa de pré-treino, o *denoising*, conforme detalhado na Seção 3.1.3. Após buscas em fóruns, foram encontradas algumas soluções anteriores que tentaram fazer o mesmo trabalho, mas foram abandonadas.

Além do passo de pré-treino usando o *dataset* BrWac para aprender o Português Brasileiro, foi gerado um vocabulário usando este mesmo *dataset*, com o objetivo de

[1] `https://huggingface.co/`

reduzir a fragmentação excessiva de palavras em muitos *tokens* e, com isso, aumentar a quantidade de palavras que o modelo consegue ingerir no novo idioma.

Ao final do pré-treino, houve a conversão do BART para o LED seguindo as ideias do código-fonte disponibilizado por Beltagy et al.[2] que necessitou ser atualizado para versão corrente da biblioteca utilizada.

Após a conversão do BART para o LED, foi realizado o ajuste fino nele usando o *dataset* RulingBR, conforme a abordagem anterior usando relatórios e votos combinados dinamicamente. Esta abordagem diminuiu bastante a Rouge medida no *dataset*, indicando a possibilidade de erro. Ao aplicar em instruções do TCU, notou-se que o resumo gerado não variava quase nada independente da instrução usada na entrada.

Como forma de entender melhor o que ocasionou um desempenho tão ruim do LED treinado, foi feita uma análise do BART pré-conversão a LED funcionando como Modelo de Linguagem para ver se ele conseguia prever corretamente o *token* mascarado (*denoising*), dado um contexto. Com exemplos simples, com *token* mascarado no início, meio ou fim e variando o assunto, o modelo mostrou que não conseguiu aprender corretamente essa tarefa, apesar de seu *loss* ter baixando de forma consistente durante o treino. Os três exemplos abaixo demonstram as predições errôneas para o Modelo de Linguagem ao tentar prever o que deveria aparecer no lugar do *token* **<mask>**:

Entrada:

Brasília é a **<mask>** do Brasil.

Top 5 saídas (ordenada da maior probabilidade para menor):

Brasília é a **provou** do Brasil.

Brasília é a **Capitania** do Brasil.

Brasília é a————————**-** do Brasil.

Brasília é a **parê** do Brasil.

Brasília é **aGIS** do Brasil.

Entrada:

O Íbis é o pior time do **<mask>**.

Top 5 saídas (ordenada da maior probabilidade para menor):

O Íbis é o pior time do**$**.

O Íbis é o pior time do **provou**.

[2] `https://github.com/allenai/longformer/blob/master/scripts/convert_bart_to_longformerencoderdecoder.py`

O Íbis é o pior time do——————-.
O Íbis é o pior time do **Operário**.
O Íbis é o pior time do**GIS**.

Entrada:
<mask> é o marido da rainha.

Top 5 saídas (ordenada da maior probabilidade para menor):
——————- é o marido da rainha.
GIS é o marido da rainha.
Exig é o marido da rainha.
$ é o marido da rainha.
polícias é o marido da rainha.

Como se pode observar pelos exemplos acima, as *top* 5 saídas mais prováveis para o *token* **<mask>** não produzem um texto sequer gramaticalmente correto ou com sentido. Portanto, essa abordagem de pré-treino usando a biblioteca Huggingface foi abandonada, preservando-se apenas o vocabulário gerado para a abordagem seguinte.

4.3.4 Pré treino do BART Fairseq no BrWac com ajuste fino do LED no RulingBR

A quarta abordagem fez exatamente o mesmo que a anterior, apenas substituindo a biblioteca Huggingface pela biblioteca Fairseq[3]. Esta biblioteca foi a que Lewis et al. [8] usaram para fazer o pré-treino do BART.

Para o pré-treino na nova biblioteca, o vocabulário gerado a partir do *dataset* BrWac para o biblioteca Huggingface foi convertido para o formato do Fairseq. Como esta biblioteca foi a mesma que realizou o pré-treino do BART, a tarefa de *denoising* já estava implementada e não foi necessário reutilizar o que havia sido desenvolvido para a biblioteca anterior.

Após o pré-treino[4], o modelo foi avaliado com o mesmo conjunto de exemplos simples utilizados na abordagem anterior, com o objetivo de avaliar se ele estava funcionando como Modelo de Linguagem corretamente. Na biblioteca Fairseq, o BART podia prever de zero a vários *tokens* para o *token* **<mask>**, conforme definido por Lewis et al. [8]. Abaixo estão os mesmos exemplos utilizados para testar o modelo da abordagem anterior, mas aplicados ao novo modelo:

[3] `https://github.com/facebookresearch/fairseq`
[4] `https://github.com/erichans/pretrain-bart-fairseq`

Entrada:

Brasília é a **<mask>** do Brasil.

Top 5 saídas (ordenada da maior probabilidade para menor):

Brasília é a **capital da República Federativa** do Brasil.
Brasília é a **capital mais populosa** do Brasil.
Brasília é a **cidade mais populosa** do Brasil.
Brasília é a **maior cidade** do Brasil.
Brasília é a **capital do estado** do Brasil.

Entrada:

O Íbis é o pior time do **<mask>**.

Top 5 saídas (ordenada da maior probabilidade para menor):

O Íbis é o pior time do **Brasil até agora**.
O Íbis é o pior time do **Brasil até hoje**.
O Íbis é o pior time do **Brasil até o momento**.
O Íbis é o pior time do **Brasil, segundo o Ibope**.
O Íbis é o pior time do **Brasil, segundo o Datafolha**.

Entrada:

<mask> é o marido da rainha.

Top 5 saídas (ordenada da maior probabilidade para menor):

- O rei é o marido da rainha.
O rei é o marido da rainha.
- O senhor é o marido da rainha.
O rei não é o marido da rainha.
- Ele é o marido da rainha.

Como pode ser observado nos exemplos acima, o Modelo de Linguagem apresentou um comportamento satisfatório e esperado para a tarefa de *denoising*, ao contrário do modelo da abordagem anterior. O ajuste fino foi beneficiado pelo comportamento correto do Modelo de Linguagem, além do entendimento do português brasileiro e, com isso, eliminou

os erros recorrentes de concordância, palavras inventadas e erro de iniciais maiúsculas em início de frases e palavras nas instruções testadas.

Após o pré-treino, o BART foi convertido do formato do Fairseq para o Huggingface e, em seguida, o BART Huggingface foi convertido para LED Huggingface para realização de ajuste fino. A Rouge medida no *dataset* RulingBR foi praticamente a mesma que a segunda abordagem, sem pré-treino (detalhada em na Subseção 4.3.2). Os resumos, apesar de terem eliminado os problemas relacionados à nossa língua, continuaram com os mesmos problemas de focar em termos muito frequentes do *dataset*.

Esta abordagem, apesar de ter o resumo similar à segunda abordagem, gera um texto com qualidade gramatical melhor.

4.3.5 Pré treino do BART Fairseq no TCU-LM com ajuste fino do LED no RulingBR

A quinta abordagem usa o Modelo de Linguagem pré-treinado para o português brasileiro e continua seu pré-treino utilizando um *dataset* jurídico, o TCU-LM (detalhes do *dataset* na Subseção 4.2.4), para que o modelo adicione ao conhecimento da nossa língua, o conteúdo jurídico presente nos textos do TCU.

A expectativa era de que, ao aprender textos jurídicos, o modelo pudesse contextualizar e entender melhor as expressões jurídicas, tais como “Agravo”, “Súmula” e “Recursos”, e assim, utilizar esses termos de maneira mais adequada, apenas quando fizessem sentido e existissem no texto de entrada.

Após a conversão do BART Fairseq, pré-treinado no TCU-LM, para BART Huggingface e posteriormente para LED Huggingface, foi efetuado o ajuste fino no RulingBR. O modelo, agora pré-treinado com conhecimento jurídico, ao contrário do ganho esperado, teve uma performance um pouco mais baixa no RulingBR, apesar dele ter absorvido conhecimento jurídico do TCU. Além de não haver melhoria, os problemas relacionados a palavras frequentes se mantiveram nesta abordagem. Como última alternativa para a abordagem utilizando a sumarização abstrativa, foi realizada a troca do *dataset* do RulingBR na abordagem seguinte.

4.3.6 Pré treino do BART Fairseq no TCU-LM com ajuste fino do LED no TCU-Summ

A sexta e última abordagem de sumarização abstrativa usa o Modelo de Linguagem pré-treinado para o Português Brasileiro com conteúdo jurídico, pré-treinado no TCU-LM, construído na abordagem anterior e substitui o *dataset* RulingBR pelo TCU-Summ,

gerado a partir de instruções do Tribunal (detalhes do *dataset* na Subseção 4.2.5). O TCU-Summ, apesar de pequeno, tem aproximadamente três vezes o tamanho de exemplos de treino do RulingBR.

Apesar de seu tamanho maior, o TCU-Summ é menos diversificado que o RulingBR e, em decorrência disso, o ajuste fino do LED trouxe problemas de memorização ainda maior para palavras frequentes. Muitos exemplos trouxeram termos comuns, como "Tomada de Contas Especial" e "auditoria", entre outros termos. Isso ocorreu, mesmo que não existissem no texto de entrada.

Além deste problema que permaneceu, outro novo foi inserido, que diz respeito à alucinação referenciando pessoas que não existiam no documento original, mas que seus nomes eram comuns em Instruções Processuais, como "Fernando" e sobrenomes como "Silva". Isso representa um problema mais grave, pois envolve o resumo imputar algo a alguém de maneira equivocada.

Com os problemas apontados nas abordagens abstrativas, esta abordagem marca o fim dos experimentos com sumarização abstrativa, pois apesar da geração de texto mais próxima do que se espera de um resumo, seus problemas ficaram bastante evidentes ao transferir o modelo treinado nos textos do STF para instruções do TCU.

4.3.7 Ajuste fino do MemSum no RulingBR

A sétima abordagem muda para forma de geração extrativa e, com isso, troca o modelo LED pelo MemSum. Este modelo foi escolhido por estar figurando entre os melhores modelos extrativos em comparativos para sumarização de documentos longos. Ele estava a frente de alguns modelos abstrativos, quando foi selecionado para este trabalho[5].

A partir deste ponto, todos os experimentos passaram a ingerir o documento inteiro ao invés de sumarizar apenas a seção EXAME TÉCNICO, pois este modelo pode ingerir até 500 sentenças (hiperparâmetro do modelo que pode ser aumentado), o que é muito maior que o limite de 16.384 *tokens* do LED. Mesmo com esta limitação de sentenças, não houve necessidade de truncar nenhuma das Instruções Processuais que foram utilizadas neste trabalho. As instruções maiores ficaram em torno de 40 páginas e couberam no modelo sem problemas. Não foi necessário treinar o modelo para suportar mais sentenças.

Uma das vantagens deste modelo é que, conforme detalhado na Seção 3.3, ele utiliza aprendizado por reforço e, por causa disso, pode otimizar direto na métrica Rouge, como sua função de recompensa, ao invés da entropia cruzada, utilizada no aprendizado supervisionado. A entropia cruzada mede a diferença entre duas distribuições de probabilidade, e sua melhoria não necessariamente indica uma melhoria na métrica Rouge. Com

[5] `https://paperswithcode.com/paper/memsum-extractive-summarization-of-long`

o aprendizado por reforço, utilizado para treinar o MemSum, esta inconsistência de treino do modelo desaparece [2].

O MemSum original usa uma técnica clássica para converter o texto de entrada em *embeddings* pré-treinados, o GloVe [37], com 200 dimensões de representação para cada *token* [16].

Para adaptar o MemSum para entender o português brasileiro, foram trocados os *embeddings* pré-treinados originais em 200 dimensões para *embeddings* pré-treinados para o português em 300 dimensões. Os *embeddings* foram obtidos no repositório do Núcleo Interinstitucional de Linguística Computacional (NILC)[6].

Como pode ser observado na Tabela 4.3, a comparação do MemSum após ajuste fino no RulingBR, com o melhor modelo LED - pré-treinado para o português e também ajustado no RulingBR (conforme discutido na Subseção 4.3.4) - demonstra um ganho em todas as métricas Rouge. Notavelmente, o MemSum, que é um modelo extrativo, superou os demais modelos abstrativos. Isso pode sugerir que os resumos do RulingBR favorecem este tipo de abordagem, e que a geração da EMENTA (saída) do *dataset* deve, copiar trechos do texto da entrada.

Tabela 4.3: Comparativo do MemSum com melhor LED no RulingBR.

	Rouge-1	**Rouge-2**	**Rouge-L**
MemSum treinado	53.0	33.8	49.9
LED treinado (somente relatório)	50.6	33.2	48.8
LED treinado (somente voto)	48.9	30.5	47.0
LED treinado (relatório + voto)	51.1	30.7	49.0

O MemSum treinado[7] foi aplicado a Instruções Processuais do TCU e trouxe fatos relevantes delas. Apesar disso, a fluidez da leitura ficou prejudicada pelo formato de geração ser extrativo. O modelo funciona como um seletor de sentenças apenas. O resumo traz informações relevantes, porém descontextualizadas. Esta forma de geração torna a leitura do resumo algo desafiador, pois não se pode entender bem do que se trata a Instrução Processual, frustrando o objetivo deste trabalho.

4.3.8 Ajuste fino do MemSum no RulingBR combinado com ajuste fino do LED na saída do MemSum

A oitava abordagem muda para forma de geração híbrida, pois o modelo extrativo, apesar de trazer fatos relevantes do texto para o resumo, gerava um conjunto de sentenças desconexas e de difícil compreensão pelo auditor sobre o conteúdo geral do texto.

[6] `http://nilc.icmc.usp.br/nilc/index.php/repositorio-de-word-embeddings-do-nilc`
[7] `https://github.com/erichans/MemSum`

Nesta abordagem, o *dataset* RulingBR foi substituído por um *dataset* gerado da seguinte forma: o MemSum foi usado para gerar o texto de entrada com as sentenças mais relevantes obtidas a partir de seu resumo, no lugar da combinação de relatório e voto, enquanto que a saída foi mantida como a EMENTA. O modelo LED foi, então, usado então para aprender a resumir a partir deste novo *dataset*, conforme a Figura 4.2.

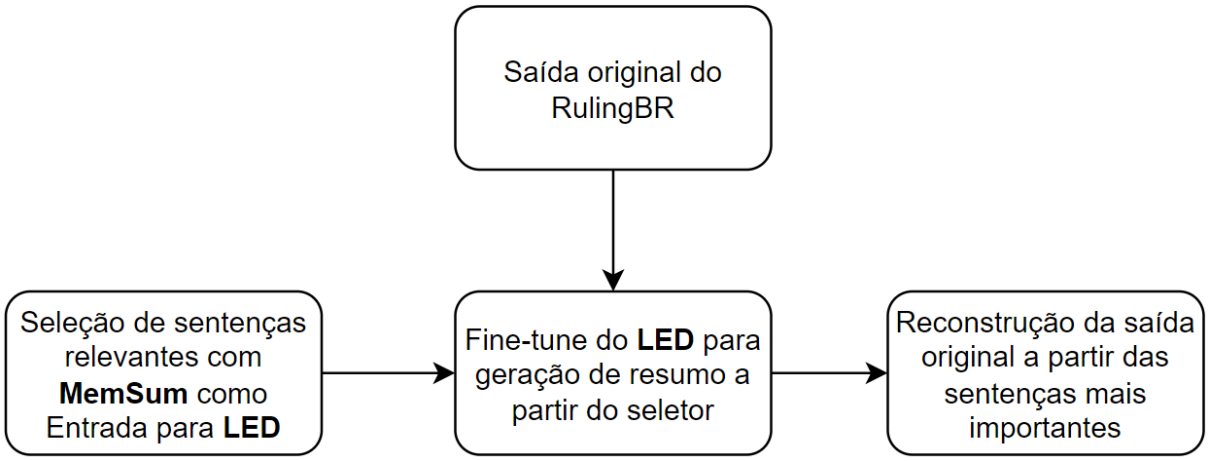

Figura 4.2: *Pipeline* para treinamento em sumarização híbrida (MemSum com LED).

A geração deste *dataset* para o treino do LED fez com que as métricas Rouge melhorassem de forma expressiva, pois o LED já recebia um conteúdo filtrado e priorizado pelo MemSum como entrada. O LED especializou-se em gerar um resumo mais focado em uma entrada muito mais reduzida, com as melhores 50 sentenças selecionadas pelo MemSum.

A Tabela 4.4 estende a Tabela 4.3, adicionando este novo modelo híbrido, que combina o MemSum com o novo LED ajustado a partir de sua seleção de sentenças.

Tabela 4.4: Comparativo do MemSum com LED e com o MemSum com novo LED.

	Rouge-1	**Rouge-2**	**Rouge-L**
MemSum treinado + LED novo treinado	62.4	46.3	60.8
MemSum treinado	53.0	33.8	49.9
LED treinado (somente relatório)	50.6	33.2	48.8
LED treinado (somente voto)	48.9	30.5	47.0
LED treinado (relatório + voto)	51.1	30.7	49.0

Apesar da melhoria na Rouge, os problemas anteriormente observados nas abordagens abstrativas ressurgiram. Dado que o texto de entrada era menor para o treinamento do LED, as palavras que já eram frequentes e significativas para o texto, tais como “Agravo”, “Súmula”, “Recursos”, entre outras, tornaram-se proporcionalmente ainda mais frequentes. O resultado final nas Instruções Processuais do TCU foi ainda pior, com uma repeti-

ção excessiva de conteúdo comum no *dataset* de treinamento. Nesta abordagem, os nomes dos ministros do STF apareceram com frequência, mesmo que não estivessem presentes nas instruções do TCU.

4.3.9 Solução final para sumarização: Ajuste fino do MemSum combinado com o ChatGPT 4k

A nona e última abordagem usa a geração híbrida, porém substitui o modelo que estava causando problemas, o LED, por um Modelo de Linguagem Grande recentemente criado, o ChatGPT com suporte a 4.096 *tokens*. Este modelo é detalhado na Seção 3.4.

Com o MemSum, os problemas relacionados à geração abstrativa, como as alucinações - as quais mencionavam pessoas inexistentes na instrução - e o foco excessivo em termos frequentes do *dataset* de treino, que não estavam presentes na instrução de entrada, foram solucionados. Entretanto, persistiram problemas relacionados à legibilidade e fluidez da leitura do texto, devido à presença de sentenças desconexas.

Para resolver este ponto, a abordagem anterior se baseou no LED e não obteve bons resultados. Com o ChatGPT (GPT 3.5 Turbo-0301 4.096 *tokens*), que tem como ponto forte a geração de textos, esse problema foi melhorado de forma significativa e sem as alucinações do LED.

O modelo ChatGPT, quando usado para gerar resumos diretamente sem a participação do MemSum, apresentou problemas, pois este modelo suporta apenas 4.096 *tokens*. É importante notar que esse limite não se aplica exclusivamente à entrada, mas à soma da entrada com o resumo gerado, o que reduz ainda mais a quantidade de texto que ele pode resumir. Para fins de comparação, apesar de suas limitações, o LED suporta 16.384 *tokens* exclusivamente para a entrada. Contudo, com documentos muito extensos, o conteúdo de entrada ainda necessita ser truncado. O ChatGPT, por sua vez, possui uma restrição de ingestão de texto ainda maior e não poderia ser utilizado diretamente sem que uma extensa parte do texto fosse suprimida, provocando, consequentemente, uma perda de informação relevante para a geração do resumo.

Além da limitação da quantidade de *tokens* do ChatGPT, seu uso direto, sem a priorização prévia das sentenças mais relevantes pelo MemSum, em alguns documentos, gerou alucinações extrínsecas [3]. O modelo extrapolou o que tinha no texto, gerando informação que não tinha sustentação no texto de entrada. Isso poderia induzir o auditor, que estivesse lendo o resumo, ao erro. O uso exclusivo do ChatGPT foi abandonado em virtude destes problemas.

Para que o ChatGPT tivesse mais contexto para geração do texto final, foram passadas as 50 sentenças mais relevantes do texto original (dentre as 500 possíveis, no máximo) para a geração do texto final.

Os dois melhores modelos, MemSum puro e MemSum combinado com o ChatGPT foram submetidos para avaliação humana e seus resultados estão detalhados na Seção 5.2.

4.4 Métricas para Avaliação dos Modelos

Para a avaliação automatizada dos modelos, foram utilizadas as métricas definidas no Capítulo 3:

1. Rouge-1/-2-L: Esta métrica foi utilizada extensivamente neste trabalho, pois é a mais utilizada para avaliação da geração de resumos. Porém sua deficiência é comparar palavras em ordens exatas, ao invés de uma avaliação mais semântica. Foi necessário adaptar o algoritmo de *stemming* para o português.

2. MoverScore: Esta métrica foi usada para fazer uma comparação mais próxima do que seria o significado semântico, utilizando o BERT [20] e WMD para cálculo de desvio entre o resumo candidato e a referência. Ela vem ganhando popularidade recentemente na área de sumarização abstrativa [2]. Entre as configurações possíveis, será usada a configuração que o autor obteve melhor resultado para sumarização. Foi necessário substituir o BERT [20] pelo BERTimbau [40] de dimensão equivalente para uso em português.

3. BERTScore: Esta métrica foi usada por fazer uma comparação do que seria mais próximo do significado semântico. Ela utiliza o $\text{RoBERTa}_{\text{LARGE}}$ [21] e uma abordagem gulosa para fazer o alinhamento entre os *tokens* dos resumos candidatos e de referência. Ela compara os *embeddings* dos *tokens* utilizando a distância de cosseno. Esta métrica possui menos variações de configuração quando comparada com o MoverScore e é, portanto, mais simples de usar. Foi necessário substituir o $\text{RoBERTa}_{\text{LARGE}}$ [21] pelo $\text{BERTimbau}_{\text{LARGE}}$ [40] para uso em português.

4.4.1 Avaliação Humana: Metodologia

Um grupo de cinco auditores participou desta parte fornecendo valiosas informações para este trabalho. Eles trabalharam em um conjunto de três Instruções Processuais públicas sorteadas, com duas variações cada:

1. **Sumarização Extrativa**: gerada pelo modelo MemSum treinado no *dataset* RulingBR, selecionando as dez sentenças mais relevantes de cada instrução

2. **Sumarização Híbrida**: gerada pelo modelo MemSum treinado no *dataset* RulingBR, selecionando as cinquenta sentenças mais relevantes de cada instrução e repassando para o ChatGPT tornar o texto mais conciso e deixar sua leitura mais agradável

O questionário (Apêndice B) usou a metodologia da antiga *Document Understanding Conferences* (DUC) e atual *Text Analysis Conference* (TAC) para avaliação da legibilidade do texto [44].

A avaliação de legibilidade se concentra em cinco aspectos [44]:

1. **Gramaticalidade**: O resumo não deve ter datas, formatação interna do sistema, erros de capitalização ou sentenças obviamente agramaticais (por exemplo, fragmentos, componentes ausentes) que dificultem a leitura do texto

2. **Não redundância**: Não deve haver repetições desnecessárias no resumo. A repetição desnecessária pode assumir a forma de frases inteiras que se repetem, ou fatos repetidos, ou o uso repetido de um substantivo ou frase nominal (por exemplo, "Bruno Dantas") quando um pronome ("ele") seria suficiente

3. **Clareza Referencial**: Deve ser fácil identificar a quem ou a que os pronomes e frases nominais no resumo estão se referindo. Se uma pessoa ou outra entidade for mencionada, deve ficar claro qual é o seu papel na história. Portanto, uma referência não seria clara se uma entidade fosse referenciada, mas sua identidade ou relação com a história permanecesse incerta.

4. **Foco**: O resumo deve ter um foco; as frases devem conter apenas informações relacionadas ao restante do resumo (sem conteúdo desconexo com o restante do resumo)

5. **Estrutura e Coerência**: O resumo deve ser bem estruturado e bem organizado. O resumo não deve ser apenas um amontoado de informações relacionadas, mas deve ser construído de frase em frase para um conteúdo coerente de informações sobre o tópico.

A seguir, apresentam-se as métricas, utilizando médias simples, para cada critério TAC [44], com o objetivo de ilustrar a avaliação de cada modelo de acordo com os cinco critérios. Os resultados são exibidos na Tabela 4.5. Para a avaliação, cinco auditores examinaram três instruções, cada uma com duas opções de resumo (extrativo e híbrido). As notas representam a média geral atribuída por todos os auditores para cada opção em cada critério. Uma avaliação é considerada positiva quando o percentual de notas é maior ou igual a quatro em cada critério.

Tabela 4.5: Avaliações humanas utilizando a TAC para o modelo extrativo e o híbrido.

	Sumarizador Extrativo	**Sumarizador Híbrido**
Gramaticalidade	1,8 (0,0% positivo)	4,8 (100,0% positivo)
Não redundância	2,9 (40,0% positivo)	4,7 (100,0% positivo)
Clareza Referencial	2,1 (20,0% positivo)	4,6 (93,3% positivo)
Foco	1,7 (13,3% positivo)	4,6 (100,0% positivo)
Estrutura e Coerência	1,5 (0,0% positivo)	4,7 (100,0% positivo)

4.5 Solução completa: Aplicação Web suportada pelo modelo de sumarização

A solução final inclui, além do modelo descrito na subseção 4.3.9, uma aplicação Web com interface gráfica semelhante ao sistema de processos do TCU, o E-TCU. Essa solução Web se integra ao modelo desenvolvido neste trabalho por meio de microsserviços utilizando *Representational State Transfer* (REST) [7].

4.5.1 Serviço Web de Sumarização: *backend* da aplicação

Atualmente, as soluções do TCU são construídas utilizando microsserviços REST, que são consumidos por aplicações Web interativas. Os serviços e as aplicações Web tem seu *deployment* feito em *Continuous Deployment* empacotados em contêineres Docker[8], orquestrados pelo Kubernetes[9] para melhor distribuição de carga e aumento de disponibilidade.

Essa abordagem arquitetural facilita a integração de novas soluções desenvolvidas utilizando tecnologias e linguagens diferentes, além de possibilitar evoluções rápidas para o usuário e, com isso, melhorar sua experiência de uso da aplicação.

A solução proposta utilizará todo o arcabouço tecnológico descrito acima e será integrada ao sistema de processos do TCU, denominado E-TCU. Com isso, ao visualizar uma Instrução Processual, o usuário terá acesso ao seu resumo gerado automaticamente. Isso permitirá que o auditor verifique rapidamente se a peça aborda o assunto de seu interesse no momento, por meio de seu resumo.

4.5.2 Aplicação Web de Sumarização: *frontend* da aplicação

A solução para o *frontend* utiliza a bibiloteca React[10] com linguagem de programação Javascript. A aplicação foi desenhada procurando trazer uma interface próxima ao que os

[8] `https://www.docker.com/`
[9] `https://kubernetes.io/`
[10] `https://react.dev/`

usuários do TCU estão acostumados a usar no E-TCU. A interface permite pesquisar por processos (Figura 4.3) e, em seguida, clicar na Instrução Processual que se deseja resumir (Figura 4.4). Qualquer peça que ele tenha acesso no E-TCU, poderá ser resumida na aplicação Web.

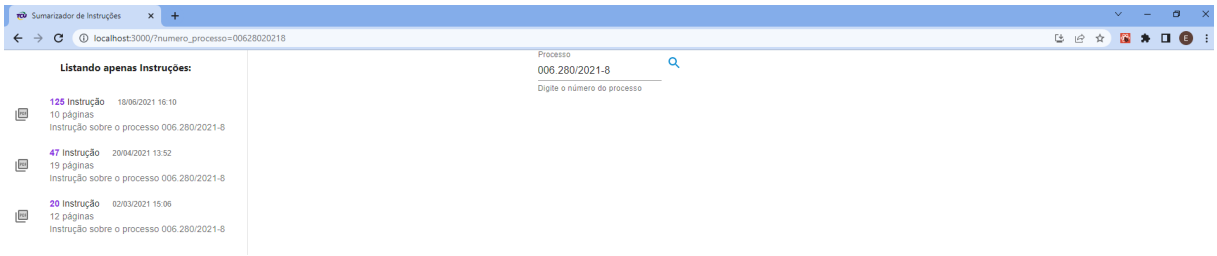

Figura 4.3: Tela de pesquisa de processos para resumo.

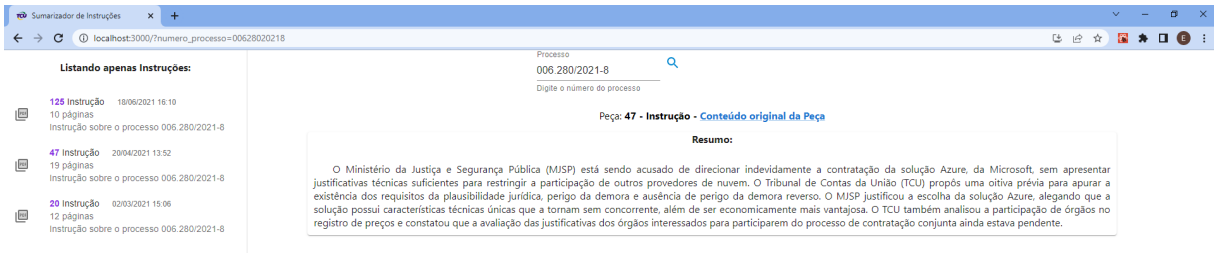

Figura 4.4: Tela de resumo.

Capítulo 5

Conclusões

Este capítulo traz as conclusões deste trabalho, a avaliação por parte dos auditores da solução final por meio do questionário (Apêndice B), além trabalhos futuros baseados no feedback dado por eles.

5.1 Aplicação Web para Sumarização

A aplicação Web desenvolvida para o TCU ficou com seu uso bem intuitivo e os auditores conseguiram usar sem necessitar treinamento ou explicação sobre seu uso, dada a similaridade de interação baseada no sistema de processos do Tribunal, o E-TCU. A melhor abordagem de sumarização, quanto à geração de texto, foi a híbrida. O sumarizador que trouxe os melhores resultados é uma combinação de um sumarizador extrativo, utilizando o MemSum com ajuste fino no *dataset* jurídico do STF, o RulingBR, seguido do ChatGPT para geração de um texto com melhor legibilidade a partir das sentenças mais relevantes selecionadas pelo extrativo. A avaliação dos auditores deixa clara a vantagem em se combinar os dois modelos, quando comparado a usar apenas o extrativo sozinho, como pode ser visto na Seção 5.2.

5.2 Avaliação humana: Questionário de avaliação dos resumos gerados

Utilizando a metodologia descrita na Subseção 4.4.1, um grupo de cinco auditores avaliou três resumos de Instruções Processuais públicas do TCU para comparação entre os dois melhores modelos finais: (i) o MemSum com ajuste fino no RulingBR, responsável pela geração de resumos extrativos, e (ii) o *pipeline* MemSum com ajuste fino no RulingBR combinado com o ChatGPT.

O resultado da avaliação apontou que o Sumarizador Híbrido teve uma avaliação muito superior que o Sumarizador Extrativo seguindo os critérios de avaliação da TAC [44]. Entre os auditores, 14,7% deles avaliaram positivamente (avaliação boa ou muito boa) o primeiro enquanto que 98,7% avaliaram positivamente o segundo.

5.3 Métricas automatizadas com maior aderência à avaliação dos auditores

Para esta avaliação, todas as métricas apresentaram valores mais elevados para o Sumarizador Híbrido, indicando corretamente que este é o melhor modelo, conforme a avaliação humana. As avaliações humanas estão agregadas utilizando média simples. O percentual de **Avaliação Humana** considerado positivo é calculado utilizando-se as notas maiores ou iguais a quatro de todas as notas em todos os critérios. A Tabela 5.1 ilustra o comparativo do sumarizadores nas Instruções Processuais do TCU.

Tabela 5.1: Comparativo das Avaliações Humanas com as Métricas Automatizadas.

	Sumarizador Extrativo	**Sumarizador Híbrido**
Avaliação Humana	2,0 (14,7% positivo)	4,7 (98,7% positivo)
Rouge-1	42,9	53,7
Rouge-2	19,7	28,7
Rouge-L	38,7	49,2
BERTScore	66,7	75,5
MoverScore	60,9	62,8

A diferença entre a opção Extrativa e a Híbrida foi grande na Avaliação Humana, enquanto que algumas métricas sinalizaram uma vantagem numérica não muito grande, como o MoverScore. Esta métrica sinalizou um aumento de apenas 1,9 pontos, enquanto que as demais mostraram um aumento mais expressivo. Portanto, desta avaliação, as métricas Rouge-1/-2/L e o BERTScore se mostraram melhores para seleção de modelos, por expressar de forma mais clara a diferença indicada pelos auditores. A métrica Rouge, por ser uma métrica de execução rápida quando comparada ao BERTScore pode ser utilizada nos modelos ainda incipientes para refinamento inicial. Quando o modelo atingir um patamar razoável nela, pode-se trocar a Rouge pelo BERTScore para refinar melhor e aproveitar seus *embeddings* para avaliar o modelo em uma maneira mais próxima do que seria a semântica do texto.

5.4 Limitações do Sumarizador Híbrido

Apesar da avaliação majoritariamente positiva do Sumarizador Híbrido, algumas questões foram apontadas no questionário de avaliação relacionadas ao conteúdo do texto, como:

1. Um dos resumos não trouxe todos os encaminhamentos propostos, podendo levar o auditor ao erro
2. Um dos auditores apontou em um dos resumos que uma pequena parte do texto poderia ser considerada ambígua
3. Um dos resumos, apontado por um auditor trouxe partes de texto que não são relevantes para o entendimento geral

Essas observações demonstram que a tarefa de sumarização é complexa e que mesmo entre avaliadores humanos há dificuldades em se avaliar o que seria um resumo correto e conciso. Não houve homogeneidade na avaliações dos auditores sobre aspectos específicos em questões abertas, mas apenas no que era relacionado aos critérios da TAC [44].

5.5 Trabalhos Futuros

A solução atual tem como objetivo fornecer uma visão geral da Instrução Processual ao auditor, mas não dispensa a leitura do documento. Ela auxilia na identificação e descarte de documentos irrelevantes para as informações que o auditor está buscando em outras Instruções Processuais que esteja trabalhando no momento.

Considerando-se o que foi apontado como limitação pelos auditores, para trabalhos futuros na sumarização de Instruções Processuais, pode-se evoluir a solução atual para realizar a sumarização baseada em *Question Answering*, pois, assim, tem-se como atender o foco específico desejado pelo usuário da aplicação. O auditor pode pedir para sumarizar e escolher pontos que ele deseja que o sumarizador foque, como materialidade, se há pedidos de cautelares, entre outras opções.

Além disso, é possível ter um resumo multidocumentos. O seletor de sentenças pode, então, escolher trechos entre os vários documentos de um mesmo processo com o objetivo de proporcionar ao auditor uma visão geral do processo.

Referências

[1] Brasil. Tribunal de Contas da União: *Regimento Interno do Tribunal de Contas da União*. https://portal.tcu.gov.br/data/files/2A/C1/CC/6A/5C66F610A6B96FE6E18818A8/BTCU_01_de_02_01_2020_Especial%20-%20Regimento_Interno.pdf, acesso em 2022-03-12. 1

[2] Alomari, Ayham, Norisma Idris, Aznul Qalid Md Sabri e Izzat Alsmadi: *Deep reinforcement and transfer learning for abstractive text summarization: A review*. Computer Speech & Language, 71:101276, 2022, ISSN 0885-2308. https://www.sciencedirect.com/science/article/pii/S0885230821000796. 3, 8, 9, 10, 17, 20, 21, 22, 30, 45, 48

[3] Ji, Ziwei, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto e Pascale Fung: *Survey of hallucination in natural language generation*. ACM Comput. Surv., 55(12), mar 2023, ISSN 0360-0300. https://doi.org/10.1145/3571730. 3, 35, 39, 47

[4] Hermann, Karl Moritz, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman e Phil Blunsom: *Teaching machines to read and comprehend*. Em Cortes, C., N. Lawrence, D. Lee, M. Sugiyama e R. Garnett (editores): *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. https://proceedings.neurips.cc/paper/2015/file/afdec7005cc9f14302cd0474fd0f3c96-Paper.pdf. 3

[5] Narayan, Shashi, Shay B. Cohen e Mirella Lapata: *Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization*. CoRR, abs/1808.08745, 2018. http://arxiv.org/abs/1808.08745. 3

[6] Beltagy, Iz, Matthew E. Peters e Arman Cohan: *Longformer: The long-document transformer*. arXiv:2004.05150, 2020. 5, 6, 7, 11, 16, 17, 18, 23

[7] Fielding, Roy Thomas: *REST: Architectural Styles and the Design of Network-based Software Architectures*. Doctoral dissertation, University of California, Irvine, 2000. http://www.ics.uci.edu/~fielding/pubs/dissertation/top.htm. 6, 50

[8] Lewis, Mike, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov e Luke Zettlemoyer: *BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension*. Em *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, páginas 7871–7880, Online, julho 2020. Association for

Computational Linguistics. https://aclanthology.org/2020.acl-main.703. 7, 10, 11, 16, 17, 18, 23, 39, 41

[9] Zhang, Jingqing, Yao Zhao, Mohammad Saleh e Peter Liu: *PEGASUS: Pre-training with extracted gap-sentences for abstractive summarization*. Em III, Hal Daumé e Aarti Singh (editores): *Proceedings of the 37th International Conference on Machine Learning*, volume 119 de *Proceedings of Machine Learning Research*, páginas 11328–11339. PMLR, 13–18 Jul 2020. https://proceedings.mlr.press/v119/zhang20ae.html. 7, 10, 11, 16, 18, 23

[10] Zaheer, Manzil, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang e Amr Ahmed: *Big bird: Transformers for longer sequences*. Em Larochelle, H., M. Ranzato, R. Hadsell, M.F. Balcan e H. Lin (editores): *Advances in Neural Information Processing Systems*, volume 33, páginas 17283–17297. Curran Associates, Inc., 2020. https://proceedings.neurips.cc/paper/2020/file/c8512d142a2d849725f31a9a7a361ab9-Paper.pdf. 7, 10, 11, 16, 23

[11] Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser e Illia Polosukhin: *Attention is all you need*. Em Guyon, I., U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan e R. Garnett (editores): *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf. 7, 9, 10, 11, 13, 14, 15

[12] Lin, Chin Yew: *ROUGE: A package for automatic evaluation of summaries*. Em *Text Summarization Branches Out*, páginas 74–81, Barcelona, Spain, julho 2004. Association for Computational Linguistics. https://aclanthology.org/W04-1013. 7, 20

[13] Zhao, Wei, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer e Steffen Eger: *Moverscore: Text generation evaluating with contextualized embeddings and earth mover distance*. CoRR, abs/1909.02622, 2019. http://arxiv.org/abs/1909.02622. 7, 20, 21, 22

[14] Zhang, Tianyi, Varsha Kishore, Felix Wu, Kilian Q. Weinberger e Yoav Artzi: *Bertscore: Evaluating text generation with BERT*. CoRR, abs/1904.09675, 2019. http://arxiv.org/abs/1904.09675. 7, 20, 22

[15] OpenAI: *Introducing chatgpt*. https://openai.com/blog/chatgpt, acesso em 2023-07-03. 7, 18, 19, 20

[16] Gu, Nianlong, Elliott Ash e Richard Hahnloser: *MemSum: Extractive summarization of long documents using multi-step episodic Markov decision processes*. Em *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, páginas 6507–6522, Dublin, Ireland, maio 2022. Association for Computational Linguistics. https://aclanthology.org/2022.acl-long.450. 7, 18, 19, 45

[17] Gupta, Som e S. K Gupta: *Abstractive summarization: An overview of the state of the art*. Expert Systems with Applications, 121:49–65, 2019, ISSN 0957-4174. https://www.sciencedirect.com/science/article/pii/S0957417418307735. 8, 20, 21

[18] Mehta, Parth: *From extractive to abstractive summarization: A journey*. Em *Proceedings of the ACL 2016 Student Research Workshop*, páginas 100–106, Berlin, Germany, agosto 2016. Association for Computational Linguistics. https://aclanthology.org/P16-3015. 8

[19] Britz, Denny, Anna Goldie, Minh Thang Luong e Quoc Le: *Massive exploration of neural machine translation architectures*. Em *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, páginas 1442–1451, Copenhagen, Denmark, setembro 2017. Association for Computational Linguistics. https://aclanthology.org/D17-1151. 9

[20] Devlin, Jacob, Ming Wei Chang, Kenton Lee e Kristina Toutanova: *BERT: Pre-training of deep bidirectional transformers for language understanding*. Em *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)*, páginas 4171–4186, Minneapolis, Minnesota, junho 2019. Association for Computational Linguistics. https://aclanthology.org/N19-1423. 10, 17, 18, 32, 48

[21] Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer e Veselin Stoyanov: *Roberta: A robustly optimized bert pretraining approach*. ArXiv, abs/1907.11692, 2019. 11, 16, 18, 22, 39, 48

[22] Oliveira, André e Anna Costa: *Plsum: Generating pt-br wikipedia by summarizing multiple websites*. Em *Anais do XVIII Encontro Nacional de Inteligência Artificial e Computacional*, páginas 751–762, Porto Alegre, RS, Brasil, 2021. SBC. https://sol.sbc.org.br/index.php/eniac/article/view/18300. 11, 31, 34, 35

[23] Raffel, Colin, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li e Peter J. Liu: *Exploring the limits of transfer learning with a unified text-to-text transformer*. Journal of Machine Learning Research, 21(140):1–67, 2020. http://jmlr.org/papers/v21/20-074.html. 11, 18, 32

[24] Carmo, Diedre, Marcos Piau, Israel Campiotti, Rodrigo Nogueira e Roberto de Alencar Lotufo: *PTT5: pretraining and validating the T5 model on brazilian portuguese data*. CoRR, abs/2008.09144, 2020. https://arxiv.org/abs/2008.09144. 11, 32

[25] Feijó, Diego de Vargas: *Summarizing legal rulings*. Tese de Doutoramento, Universidade Federal do Rio Grande do Sul, 2021. https://www.lume.ufrgs.br/handle/10183/230669. 11, 36

[26] Hochreiter, Sepp e Jürgen Schmidhuber: *Long Short-Term Memory*. Neural Computation, 9(8):1735–1780, novembro 1997, ISSN 0899-7667. https://doi.org/10.1162/neco.1997.9.8.1735. 13

[27] Cho, Kyunghyun, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk e Yoshua Bengio: *Learning phrase representations using RNN encoder–decoder for statistical machine translation*. Em *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, páginas 1724–1734, Doha, Qatar, outubro 2014. Association for Computational Linguistics. https://aclanthology.org/D14-1179. 13

[28] He, Kaiming, Xiangyu Zhang, Shaoqing Ren e Jian Sun: *Deep residual learning for image recognition*. Em *Proceedings of the IEEE conference on computer vision and pattern recognition*, páginas 770–778, 2016. 13

[29] Ba, Jimmy Lei, Jamie Ryan Kiros e Geoffrey E Hinton: *Layer normalization*. arXiv preprint arXiv:1607.06450, 2016. 13

[30] Tay, Yi, Mostafa Dehghani, Dara Bahri e Donald Metzler: *Efficient transformers: A survey*. ACM Comput. Surv., 55(6), dec 2022, ISSN 0360-0300. https://doi.org/10.1145/3530811. 15, 16

[31] Clark, Kevin, Urvashi Khandelwal, Omer Levy e Christopher D. Manning: *What does BERT look at? an analysis of BERT's attention*. Em *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, páginas 276–286, Florence, Italy, agosto 2019. Association for Computational Linguistics. https://aclanthology.org/W19-4828. 14

[32] Google: *Constructing transformers for longer sequences with sparse attention methods*. https://ai.googleblog.com/2021/03/constructing-transformers-for-longer.html, acesso em 2023-08-07. 16

[33] Cohan, Arman, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang e Nazli Goharian: *A discourse-aware attention model for abstractive summarization of long documents*. Em *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, páginas 615–621, New Orleans, Louisiana, junho 2018. Association for Computational Linguistics. https://aclanthology.org/N18-2097. 16

[34] Sharma, Eva, Chen Li e Lu Wang: *BIGPATENT: A large-scale dataset for abstractive and coherent summarization*. Em *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, páginas 2204–2213, Florence, Italy, julho 2019. Association for Computational Linguistics. https://aclanthology.org/P19-1212. 16

[35] Brown, Tom, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell *et al.*: *Language models are few-shot learners*. Advances in neural information processing systems, 33:1877–1901, 2020. 18, 19

[36] OpenAI: *Function calling and other api updates*. https://openai.com/blog/function-calling-and-other-api-updates, acesso em 2023-07-03. 18, 20

[37] Pennington, Jeffrey, Richard Socher e Christopher Manning: *GloVe: Global vectors for word representation*. Em *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, páginas 1532–1543, Doha, Qatar, outubro 2014. Association for Computational Linguistics. https://aclanthology.org/D14-1162. 18, 45

[38] Liu, Yang e Mirella Lapata: *Hierarchical transformers for multi-document summarization*. Em *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, páginas 5070–5081, Florence, Italy, julho 2019. Association for Computational Linguistics. https://aclanthology.org/P19-1500. 19

[39] Gatt, Albert e Emiel Krahmer: *Survey of the state of the art in natural language generation: Core tasks, applications and evaluation*. J. Artif. Int. Res., 61(1):65–170, jan 2018, ISSN 1076-9757. https://dl.acm.org/doi/10.5555/3241691.3241693. 20, 21

[40] Souza, Fábio, Rodrigo Nogueira e Roberto Lotufo: *Bertimbau: Pretrained bert models for brazilian portuguese*. Em Cerri, Ricardo e Ronaldo C. Prati (editores): *Intelligent Systems*, páginas 403–417, Cham, 2020. Springer International Publishing, ISBN 978-3-030-61377-8. 22, 32, 39, 48

[41] Liu, Peter J., Mohammad Saleh, Etienne Pot, Ben Goodrich, Ryan Sepassi, Lukasz Kaiser e Noam Shazeer: *Generating wikipedia by summarizing long sequences*. CoRR, abs/1801.10198, 2018. http://arxiv.org/abs/1801.10198. 31

[42] Wagner Filho, Jorge A., Rodrigo Wilkens, Marco Idiart e Aline Villavicencio: *The brWaC corpus: A new open resource for Brazilian Portuguese*. Em *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, maio 2018. European Language Resources Association (ELRA). https://aclanthology.org/L18-1686. 31, 32

[43] Vargas Feijó, Diego de e Viviane Pereira Moreira: *Rulingbr: A summarization dataset for legal texts*. Em Villavicencio, Aline, Viviane Moreira, Alberto Abad, Helena Caseli, Pablo Gamallo, Carlos Ramisch, Hugo Gonçalo Oliveira e Gustavo Henrique Paetzold (editores): *Computational Processing of the Portuguese Language*, páginas 255–264, Cham, 2018. Springer International Publishing, ISBN 978-3-319-99722-3. 31

[44] Dang, Hoa: *Overview of duc 2005*. página 48, janeiro 2006. https://duc.nist.gov/pubs/2005papers/OVERVIEW05.pdf. 49, 53, 54

Apêndice A

Questionário sobre necessidades de sumarização

Sumarização de documentos usando Inteligência Artificial

Público alvo: auditores que trabalhem com instrução processual e/ou fiscalizações.

Este questionário é parte do entendimento dos benefícios em se gerar resumos (sumarização) de textos de forma automática utilizando técnicas de Inteligência Artificial no TCU. É parte do meu mestrado desenvolver a solução inicial para o problema.

Conto com suas informações para refinar melhor o problema e atuar gerando o maior benefício para o Tribunal como um todo e reduzir o tempo gasto na leitura de documentos que poderiam já trazer um resumo gerado por Inteligência Artifical.

Os dados serão usados apenas de maneira agregada com o objetivo de ter uma visão melhor das necessidades dos Auditores do TCU.

*Obrigatório

1. Qual seu e-mail do TCU? (Caso necessário para dirimir dúvidas pontuais. Apenas se autorizado por você ao fim do questionário) *

2. Qual secretaria você está lotado? *

3. Quantos anos de Tribunal você tem? *

Marcar apenas uma oval.

- ◯ Até 3 anos
- ◯ De 3 a 5 anos
- ◯ Mais de 5 anos

4. Você participou de alguma fiscalização (auditorias, levantamentos, inspeções, acompanhamentos, monitoramentos) nos últimos 3 anos? *

Marcar apenas uma oval.

- ◯ Sim
- ◯ Não

5. Quais tipos de trabalhos você mais atua? *

Marque todas que se aplicam.

- ☐ Auditorias
- ☐ Levantamentos
- ☐ Acompanhamentos
- ☐ Monitoramentos
- ☐ Outro: _____

6. Como você avalia o tempo gasto para ler **Peças de processos do TCU relacionadas ao assunto** durante uma fiscalização ou instrução (Seja pelo volume deles ou pelo tamanho em geral)? (Isso será usado para escolher os documentos que teremos mais ganho ao resumi-los) *

Marcar apenas uma oval.

	1	2	3	4	5	
Pouco tempo gasto	◯	◯	◯	◯	◯	Muito tempo gasto

7. Como você avalia o tempo gasto para ler **Notícias sobre o assunto** durante uma fiscalização ou instrução (Seja pelo volume deles ou pelo tamanho em geral)? (Isso será usado para escolher os documentos que teremos mais ganho ao resumi-los) *

Marcar apenas uma oval.

	1	2	3	4	5	
Pouco tempo gasto	◯	◯	◯	◯	◯	Muito tempo gasto

8. Como você avalia o tempo gasto para ler **Normativos sobre o assunto** durante uma fiscalização ou instrução (Seja pelo volume deles ou pelo tamanho em geral)? (Isso será usado para escolher os documentos que teremos mais ganho ao resumi-los) *

Marcar apenas uma oval.

	1	2	3	4	5	
Pouco tempo gasto	◯	◯	◯	◯	◯	Muito tempo gasto

9. Como você avalia o tempo gasto para ler **Documentos recebidos da Unidade Jurisdicionada, responsável ou representante legal (Ex.: Resposta à comunicação)** durante uma fiscalização ou instrução (Seja pelo volume deles ou pelo tamanho em geral)? *

Marcar apenas uma oval.

	1	2	3	4	5	
Pouco tempo gasto	◯	◯	◯	◯	◯	Muito tempo gasto

10. Caso existam outros tipos de documento que consumam bastante seu tempo (a partir de 3 na escala das perguntas anteriores), poderia descrevê-los abaixo? O volume é elevado? São documentos grandes? Ambos os casos?

_____

_____

_____

_____

_____

Leitura de documentos de processos anteriores no TCU

11. Entre os processos do TCU relacionados ao assunto, o quanto você lê de outras **Instruções**? *

Marcar apenas uma oval.

	1	2	3	4	5	
Pouca leitura	◯	◯	◯	◯	◯	Muita leitura

12. Entre os processos do TCU relacionados ao assunto, o quanto você lê de outros **Relatórios**? *

Marcar apenas uma oval.

	1	2	3	4	5	
Pouca leitura	◯	◯	◯	◯	◯	Muita leitura

13. Entre os processos do TCU relacionados ao assunto, o quanto você lê de outros **Votos**? *

Marcar apenas uma oval.

	1	2	3	4	5	
Pouca leitura	◯	◯	◯	◯	◯	Muita leitura

14. Entre os processos do TCU relacionados ao assunto, o quanto você lê de outros **Acórdãos**? *

Marcar apenas uma oval.

	1	2	3	4	5	
Pouca leitura	◯	◯	◯	◯	◯	Muita leitura

15. Qual dos documentos acima consome mais tempo seu e haveria benefício ao se gerar um resumo automático para você? *

Marque todas que se aplicam.

- [ ] Instruções
- [ ] Relatórios
- [ ] Votos
- [ ] Acórdãos
- [ ] Resposta à comunicação
- [ ] Outro: _____

16. Qual o tamanho médio desses documentos (aproximadamente)? *

Marcar apenas uma oval.

- ◯ 1 a 10 páginas
- ◯ 11 a 30 páginas
- ◯ 31 a 50 páginas
- ◯ 51 a 100 páginas
- ◯ Mais de 100 páginas

17. O que mais consome tempo para leitura desses conteúdos? *

Marcar apenas uma oval.

- ◯ Quantidade alta de documentos para ler
- ◯ Tamanho de cada documento para ler
- ◯ Ambos

18. Geração automática de resumos de seções específicas de documentos de fiscalização ou instruções são relevantes para otimizar seu tempo (Ex.: Questões de auditoria, limitações, achados, etc.)? *

Marcar apenas uma oval.

- Sim
- Não

19. Caso positivo, quais seções?

20. Caso tenha alguma observação a fazer ou informação a complementar, pode usar o espaço abaixo:

21. Posso entrar em contato com você por e-mail/Teams, caso precise tirar algumas dúvidas rápidas sobre as respostas do questionário, de acordo com sua disponibilidade? *

Marcar apenas uma oval.

- Sim
- Não

Apêndice B

Questionário de Avaliação dos Resumos

Avaliação de resumos de instruções usando Inteligência Artificial

Este questionário é parte da avaliação dos resumos de textos gerados de forma automática utilizando técnicas de Inteligência Artificial no TCU modernas. É parte do meu trablho avaliar o grau de qualidade do resumo gerado de forma objetiva.

Conto com suas informações para refinar melhor o problema e atuar gerando o maior benefício para o Tribunal como um todo e reduzir o tempo gasto na leitura de documentos que poderiam já trazer um resumo gerado por Inteligência Artifical.

Os dados serão utilizados para entender melhor os pontos fracos e fortes dos resumos gerados utilizando o que existe de estado-da-arte em IA e direcionar trabalhos futuros na área.

Serão apresentadas 3 instruções (com link para o e-tcu) e 2 opções de resumo para cada.

Obrigado pela participação!

* Indica uma pergunta obrigatória

1. E-mail *

Para o documento Instrução 1 (clique no link da instrução para ler o conteúdo e avaliar as opções de resumo dele) serão apresentadas opções de resumo geradas e perguntas sobre a qualidade do resumo gerado. As perguntas são idênticas para cada opção para que se possa comparar os resultados de forma objetiva

Resumo da Instrução 1 usando **Método de Seleção de Sentenças Mais Relevantes**:

Complementando a análise sobre o item, a Siurb/SP afirma que eventual somatório de quantitativos dos serviços integrantes dos atestados seria descabido para o caso concreto, tendo em vista a necessidade de aferir a capacidade operacional da empresa.

Serviço de instalação de bombeamento e implantação de Sistema de Abastecimento de Água com chafariz de 5000 L, com energização em sistema autônomo de geração fotovoltaica, na quantidade mínima de 30% (trinta por cento) do montante de cada lote relacionado ao sistema autônomo de geração fotovoltaica (no caso do interesse de participação em mais de um).

TC 007.084/2022-6 Apenso: não há Tipo: Denúncia Unidade Jurisdicionada: Fundação Nacional de Saúde (CNPJ: 26.989.350/0001-16 e UASG: 255000) Denunciante: identidade preservada (art. 55 da Lei 8.443/1992) Procurador: não há Interessado em sustentação oral: não há Proposta: preliminar (conhecer/conceder cautelar/realizar oitivas e diligência) INTRODUÇÃO

Da impugnação do edital do edital do Pregão Eletrônico-SRP 3/2022 I.1.

Serviço de bombeamento com análises físico-químicas-bacteriológicas em poço tubular profundo, na quantidade mínima de 30% do montante de cada lote (no caso do interesse de participação em mais de um);

Já o art. 11 da Lei 10.520/2002 permite a utilização de pregão com registro de preços para bens e serviços comuns.

Da adoção do sistema registro de preços (SRP) para a contratação

Da falta de especificação adequada dos equipamentos licitados

Acórdãos 2.150/2008, 2.882/2008, 1.237/2008, 1.636/2007, 2.359/2007 e 2.019/2013, todos do Plenário).

Assim, propõe-se, com fundamento no art. 276, do Regimento Interno do TCU, deferir o pedido de concessão de medida cautelar, sem oitiva prévia, e determinar que a Fundação Nacional de Saúde (Funasa) suspenda o Pregão Eletrônico SRP 3/2022 até que o Tribunal delibere sobre o mérito da matéria.

2. **Gramaticalidade**: O resumo não deve ter datas, formatação interna do sistema, erros de capitalização ou sentenças obviamente agramaticais (por exemplo, fragmentos, componentes ausentes) que dificultem a leitura do texto. *

Marcar apenas uma oval.

Muito ruim

1
2
3
4
5

Muito bom

3. **Não redundância**: Não deve haver repetições desnecessárias no resumo. A repetição desnecessária pode assumir a forma de frases inteiras que se repetem, ou fatos repetidos, ou o uso repetido de um substantivo ou frase nominal (por exemplo, “Bruno Dantas”) quando um pronome (“ele”) seria suficiente. *

Marcar apenas uma oval.

Muito ruim

1
2
3
4
5

Muito bom

4. **Clareza Referencial**: Deve ser fácil identificar a quem ou a que os pronomes e frases nominais no resumo estão se referindo. Se uma pessoa ou outra entidade for mencionada, deve ficar claro qual é o seu papel na história. Portanto, uma referência não seria clara se uma entidade fosse referenciada, mas sua identidade ou relação com a história permanecesse incerta. *

Marcar apenas uma oval.

- Muito ruim
- 1
- 2
- 3
- 4
- 5
- Muito bom

5. **Foco**: O resumo deve ter um foco; as frases devem conter apenas informações relacionadas ao restante do resumo (sem conteúdo desconexo com o restante do resumo) *

Marcar apenas uma oval.

- Muito ruim
- 1
- 2
- 3
- 4
- 5
- Muito bom

6. **Estrutura e Coerência**: O resumo deve ser bem estruturado e bem organizado. O resumo não deve ser apenas um amontoado de informações relacionadas, mas deve ser construído de frase em frase para um conteúdo coerente de informações sobre o tópico. *

Marcar apenas uma oval.

Muito ruim
1
2
3
4
5
Muito bom

7. Houve algum ato ou fato jurídico gerado no resumo que contradissesse ou extrapolasse o que estava no texto? Caso afirmativo, por favor, descreva abaixo onde ocorreram e qual ponto de contradição e/ou extrapolação que a instrução não dá suporte ao que foi gerado no resumo. *

Resumo da Instrução 1 usando **Método de Modelo Generativo Priorizado por Relevância**:
O texto trata de uma denúncia de possíveis irregularidades em um processo licitatório da Fundação Nacional de Saúde (Funasa) para a contratação de empresa especializada na realização de serviço de bombeamento e instalação de sistemas simplificados de abastecimento de água em poços tubulares profundos em áreas rurais de vários estados brasileiros. A denúncia apresenta impugnações ao edital do pregão eletrônico e questiona a falta de informações sobre a localização e situação dos poços, a não divisão dos serviços licitados e a falta de especificação adequada dos equipamentos. O texto também destaca que o sistema de registro de preços foi utilizado no processo licitatório. O Tribunal de Contas da União (TCU) irá analisar o caso com a possibilidade de adoção de medida cautelar.

8. **Gramaticalidade**: O resumo não deve ter datas, formatação interna do sistema, erros de capitalização ou sentenças obviamente agramaticais (por exemplo, fragmentos, componentes ausentes) que dificultem a leitura do texto. *

Marcar apenas uma oval.

Muito ruim
1
2
3
4
5
Muito bom

9. **Não redundância**: Não deve haver repetições desnecessárias no resumo. A repetição desnecessária pode assumir a forma de frases inteiras que se repetem, ou fatos repetidos, ou o uso repetido de um substantivo ou frase nominal (por exemplo, “Bruno Dantas”) quando um pronome (“ele”) seria suficiente. *

Marcar apenas uma oval.

Muito ruim

1
2
3
4
5

Muito bom

10. **Clareza Referencial**: Deve ser fácil identificar a quem ou a que os pronomes e frases nominais no resumo estão se referindo. Se uma pessoa ou outra entidade for mencionada, deve ficar claro qual é o seu papel na história. Portanto, uma referência não seria clara se uma entidade fosse referenciada, mas sua identidade ou relação com a história permanecesse incerta. *

Marcar apenas uma oval.

Muito ruim

1
2
3
4
5

Muito bom

11. **Foco**: O resumo deve ter um foco; as frases devem conter apenas informações relacionadas ao restante do resumo (sem conteúdo desconexo com o restante do resumo) *

Marcar apenas uma oval.

Muito ruim
1
2
3
4
5
Muito bom

12. **Estrutura e Coerência**: O resumo deve ser bem estruturado e bem organizado. O resumo não deve ser apenas um amontoado de informações relacionadas, mas deve ser construído de frase em frase para um conteúdo coerente de informações sobre o tópico. *

Marcar apenas uma oval.

Muito ruim
1
2
3
4
5
Muito bom

13. Houve algum ato ou fato jurídico gerado no resumo que contradissesse ou extrapolasse o que estava no texto? Caso afirmativo, por favor, descreva abaixo onde ocorreram e qual ponto de contradição e/ou extrapolação que a instrução não dá suporte ao que foi gerado no resumo. *

14. Quanto tempo total você levou para ler toda instrução e avaliar cada opção de resumo gerada? *

15. Qual a melhor opção de resumo gerado? (Você pode utilizar o botão voltar do formulário para revê-los) *

Marcar apenas uma oval.

- Método de Seleção de Sentenças Mais Relevantes
- Método de Modelo Generativo Priorizado por Relevância

16. (Opcional) Caso haja alguma falha no melhor resumo apontado por você, por favor, detalhe o que faltou nele ou o que ele colocou a mais de forma indevida.

Para o documento Instrução 2 (clique no link da instrução para ler o conteúdo e avaliar as opções de resumo dele) serão apresentadas opções de resumo geradas e perguntas sobre a qualidade do resumo gerado. As perguntas são idênticas para cada opção para que se possa comparar os resultados de forma objetiva

Resumo da Instrução 2 usando **Método de Seleção de Sentenças Mais Relevantes**:

Quanto à plausibilidade jurídica do pedido (fumus boni iuris), a análise constante dos de Fiscalização de Tecnologia da Informação parágrafos 27 a 44 da presente instrução revela a inexistência de indícios de irregularidade suficientes para caracterizar tal pressuposto, no que tange ao direcionamento da contratação da solução Azure por parte do MJSP.

encaminhar cópia da decisão que vier a ser adotada e da presente instrução ao Ministério da Justiça e Segurança Pública, à Polícia Rodoviária Federal, ao Conselho Administrativo de Defesa Econômica, à Polícia Federal, à Fundação Nacional do Índio, bem como à empresa Lanlink Soluções e Comercializações em Informática S/A (CNPJ 19.877.285/0002-52), a fim de subsidiar as manifestações a serem apresentadas.

Por outro lado, no que tange à aquisição a ser realizada pelos órgãos partícipes da ata de registro de preços, considerando a inexistência de informações nos autos que permitam se confirmar pela presença do perigo da demora reverso de eventual medida cautelar adotada por esta Corte, bem como a de Fiscalização de Tecnologia da Informação possiblidade de que sejam apresentadas informações e justificativas que ajudem a elucidar o indício de irregularidade constatado, propõe-se a oitiva prévia dos órgãos partícipes da ARP e da empresa Lanlink, bem como diligência do MJSP (parágrafos 74 a 76).

o MJSP mantém em vigência o Contrato 28/2018 de serviços de computação em nuvem Azure, cujo encerramento está previsto para 27/12/2021.

de Fiscalização de Tecnologia da Informação Análise

TC 006.280/2021-8 de Fiscalização de Tecnologia da Informação Tipo: Denúncia (com pedido de medida cautelar) Unidade jurisdicionada: Ministério da Justiça e Segurança Pública (MJSP) Denunciante: identidade preservada (Lei 8.443/1992, art. 55) Advogado ou Procurador: não há Interessado em sustentação oral: não há Proposta: Oitiva, diligência.

Consoante o art. 276 do Regimento Interno/TCU, o Relator poderá, em caso de urgência, de fundado receio de grave lesão ao Erário, ao interesse público, ou de risco de ineficácia da decisão de mérito, de ofício ou mediante provocação, adotar medida cautelar, determinando a suspensão do procedimento impugnado, até que o Tribunal julgue o mérito da questão.

realizar diligência, com fundamento no art. 157 do RI/TCU, ao Ministério da Justiça e Segurança Pública para que, no prazo de cinco dias úteis, encaminhe ao Tribunal as análises que eventualmente já tenham sido realizadas quanto às justificativas dos órgãos participantes do Pregão Eletrônico 1/2021, quanto à compatibilidade de seus

ETP e de outros documentos de planejamento da contratação com o TR ou PB do MJ, com base no § 1º, inciso III, e § 3º, do art. 9º, da Instrução Normativa - 1/2019;

Ainda no âmbito da instrução anterior, expôs-se que o exame das irregularidades denunciadas dos demais elementos presentes nos autos, notadamente o edital do certame e o ETP, a indicação da referida solução de nuvem encontrava-se desprovida de justificativas suficientes que fundamentassem o direcionamento do objeto ao produto da empresa Microsoft, (peça 12, p. 9).

Nessa esteira, em consulta aos autos do Processo SEI 08006.000110/2020-85, foram identificados os comprovantes de manifestação de interesse dos órgãos participantes do registro de preços (peça 45, p. 1-8), em atendimento, portanto, ao art. 6º do Decreto 7.892/2013.

17. **Gramaticalidade**: O resumo não deve ter datas, formatação interna do sistema, erros de capitalização ou sentenças obviamente agramaticais (por exemplo, fragmentos, componentes ausentes) que dificultem a leitura do texto. *

Marcar apenas uma oval.

	Muito ruim
1	
2	
3	
4	
5	
	Muito bom

18. **Não redundância**: Não deve haver repetições desnecessárias no resumo. A repetição desnecessária pode assumir a forma de frases inteiras que se repetem, ou fatos repetidos, ou o uso repetido de um substantivo ou frase nominal (por exemplo, “Bruno Dantas”) quando um pronome (“ele”) seria suficiente. *

Marcar apenas uma oval.

	Muito ruim
1	
2	
3	
4	
5	
	Muito bom

19. **Clareza Referencial**: Deve ser fácil identificar a quem ou a que os pronomes e frases nominais no resumo estão se referindo. Se uma pessoa ou outra entidade for mencionada, deve ficar claro qual é o seu papel na história. Portanto, uma referência não seria clara se uma entidade fosse referenciada, mas sua identidade ou relação com a história permanecesse incerta. *

Marcar apenas uma oval.

- Muito ruim
- 1
- 2
- 3
- 4
- 5
- Muito bom

20. **Foco**: O resumo deve ter um foco; as frases devem conter apenas informações relacionadas ao restante do resumo (sem conteúdo desconexo com o restante do resumo) *

Marcar apenas uma oval.

- Muito ruim
- 1
- 2
- 3
- 4
- 5
- Muito bom

21. **Estrutura e Coerência**: O resumo deve ser bem estruturado e bem organizado. O resumo não deve ser apenas um amontoado de informações relacionadas, mas deve ser construído de frase em frase para um conteúdo coerente de informações sobre o tópico. *

Marcar apenas uma oval.

Muito ruim

1
2
3
4
5

Muito bom

22. Houve algum ato ou fato jurídico gerado no resumo que contradissesse ou extrapolasse o que estava no texto? Caso afirmativo, por favor, descreva abaixo onde ocorreram e qual ponto de contradição e/ou extrapolação que a instrução não dá suporte ao que foi gerado no resumo. *

Resumo da Instrução 2 usando **Método de Modelo Generativo Priorizado por Relevância**:
O Ministério da Justiça e Segurança Pública (MJSP) está sendo investigado pelo Tribunal de Contas da União (TCU) em relação a uma denúncia de direcionamento de um certame para a contratação da solução Azure, da Microsoft, sem justificativas técnicas suficientes. O TCU propôs a realização de uma oitiva prévia para melhor apuração dos requisitos da plausibilidade jurídica e do perigo da demora, além da ausência de perigo da demora reverso, antes de decidir sobre o pedido de medida cautelar. Após as informações prestadas pelo MJSP, o TCU entendeu que a escolha da solução Azure encontra-se técnica e razoavelmente justificada, em atendimento ao art. 7º, § 5º da Lei 8.666/1994, que permite a indicação de marca em licitações públicas, desde que justificada por razões técnicas. O TCU também analisou o procedimento de participação de órgãos no registro de preços e constatou que não foi realizada análise de compatibilidade do Estudo Técnico Preliminar (ETP) e outros documentos de planejamento da contratação do órgão interessado na participação com o Termo de Referência (TR) ou Projeto Básico (PB) do órgão gerenciador, conforme preceitua a Instrução Normativa - SGD 1/2019.

23. **Gramaticalidade**: O resumo não deve ter datas, formatação interna do sistema, erros de capitalização ou sentenças obviamente agramaticais (por exemplo, fragmentos, componentes ausentes) que dificultem a leitura do texto. *

Marcar apenas uma oval.

Muito ruim

1
2
3
4
5

Muito bom

24. **Não redundância**: Não deve haver repetições desnecessárias no resumo. A repetição desnecessária pode assumir a forma de frases inteiras que se repetem, ou fatos repetidos, ou o uso repetido de um substantivo ou frase nominal (por exemplo, “Bruno Dantas”) quando um pronome (“ele”) seria suficiente. *

Marcar apenas uma oval.

Muito ruim

1
2
3
4
5

Muito bom

25. **Clareza Referencial**: Deve ser fácil identificar a quem ou a que os pronomes e frases nominais no resumo estão se referindo. Se uma pessoa ou outra entidade for mencionada, deve ficar claro qual é o seu papel na história. Portanto, uma referência não seria clara se uma entidade fosse referenciada, mas sua identidade ou relação com a história permanecesse incerta. *

Marcar apenas uma oval.

Muito ruim

1
2
3
4
5

Muito bom

26. **Foco**: O resumo deve ter um foco; as frases devem conter apenas informações relacionadas ao restante do resumo (sem conteúdo desconexo com o restante do resumo) *

Marcar apenas uma oval.

Muito ruim
1
2
3
4
5
Muito bom

27. **Estrutura e Coerência**: O resumo deve ser bem estruturado e bem organizado. O resumo não deve ser apenas um amontoado de informações relacionadas, mas deve ser construído de frase em frase para um conteúdo coerente de informações sobre o tópico. *

Marcar apenas uma oval.

Muito ruim
1
2
3
4
5
Muito bom

28. Houve algum ato ou fato jurídico gerado no resumo que contradissesse ou extrapolasse o que estava no texto? Caso afirmativo, por favor, descreva abaixo onde ocorreram e qual ponto de contradição e/ou extrapolação que a instrução não dá suporte ao que foi gerado no resumo. *

29. Quanto tempo total você levou para ler toda instrução e avaliar cada opção de resumo gerada? *

30. Qual a melhor opção de resumo gerado? (Você pode utilizar o botão voltar do formulário para revê-los) *

Marcar apenas uma oval.

- Método de Seleção de Sentenças Mais Relevantes
- Método de Modelo Generativo Priorizado por Relevância

31. (Opcional) Caso haja alguma falha no melhor resumo apontado por você, por favor, detalhe o que faltou nele ou o que ele colocou a mais de forma indevida.

Para o documento Instrução 3 (clique no link da instrução para ler o conteúdo e avaliar as opções de resumo dele) serão apresentadas opções de resumo geradas e perguntas sobre a qualidade do resumo gerado. As perguntas são idênticas para cada opção para que se possa comparar os resultados de forma objetiva

Resumo da Instrução 3 usando **Método de Seleção de Sentenças Mais Relevantes**:

Hoje a gente tem o laudo pericial que classifica a operação feita de extração indevida de dados.

Arquitetura básica de uma aplicação web tradicional;

Ato originário: Acórdão 1.413/2021-TCU Plenário (peça 11).

QUESTÃO DE AUDITORIA 01 (QST-01): Processo de trabalho + Local [Quem?]

Ato de designação: Portaria de Fiscalização-Sefti 336/2021 (peça 16).

As principais lógicas implementadas são a soma dos fatores de riscos, o cálculo do escore de gravidade da Covid-19 e a lógica de ramificação que apresenta diagnóstico e sugestão de conduta (todos detalhados mais abaixo);

3 Funcionamento do TrateCov e medicamentos recomendados

RELATÓRIO DE INSPEÇÃO TC 015.749/2021-5 Fiscalização 119/2021 Relator: Ministro Vital do Rêgo.

RISCOS MAPEADOS (Código sequencial para cada risco do relatório: RIS-01, RIS-02, ..., RIS-XX) DETALHAMENTO POSSÍVEIS ACHADOS POSSÍVEIS INFORMAÇÕES FONTES DE CRITÉRIOS DOS (PEDIDO CPI) EVIDÊNCIAS REQUERIDAS INFORMAÇÃO PROCEDIMENTOS c) Arquitetura do TrateCov c1) o que é o TrateCov c2) o que pode ser considerado código fonte do TrateCov (o que observar para avaliar como a aplicação funciona) I2 - “Verificar se na □ Relatório de Não há (questão Solicitar e receber a) Informações que Ministério da Saúde versão originária desse inspeção do arquivo com descritiva) arquivos de projeto, com devem constar do aplicativo havia previsão metadados do projeto cuidados de garantia de relatório de inspeção do de tratamento precoce e (papel de trabalho) integridade arquivo com metadados: quais seriam os □ Manual do Checar integridade do a.1) Descrição dos medicamentos procedimentos de sistema (peça 37 do TC arquivo recebido (hash) recomendados (e.g.

RISCOS MAPEADOS (Código sequencial para cada risco do relatório: RIS-01, RIS-02, ..., RIS-XX) DETALHAMENTO POSSÍVEIS ACHADOS POSSÍVEIS INFORMAÇÕES FONTES DE CRITÉRIOS DOS (PEDIDO CPI) EVIDÊNCIAS REQUERIDAS INFORMAÇÃO PROCEDIMENTOS TODOS □ Ofício de requisição (697/2021) □ Ofício de resposta do MS e anexos I1 - “Avaliar a □ Relatório de Não há (questão Descrever a arquitetura de a) Funcionamento uma Ofício de resposta do MS arquitetura do aplicativo revisão analítica da descritiva) uma aplicação web aplicação web Internet (artigos, blogs e TrateCov” arquitetura (produzido tradicional tradicional postagens em formato pela equipe) a1) Conceitos: aplicação digital sobre arquiteturas

de Documentos Descrever a arquitetura de Web x App x... aplicações web) 18 QUESTÃO DE AUDITORIA 01 (QST-01): Processo de trabalho + Local [Quem?]

32. **Gramaticalidade**: O resumo não deve ter datas, formatação interna do sistema, erros de capitalização ou sentenças obviamente agramaticais (por exemplo, fragmentos, componentes ausentes) que dificultem a leitura do texto. *

Marcar apenas uma oval.

- Muito ruim
- 1
- 2
- 3
- 4
- 5
- Muito bom

33. **Não redundância**: Não deve haver repetições desnecessárias no resumo. A repetição desnecessária pode assumir a forma de frases inteiras que se repetem, ou fatos repetidos, ou o uso repetido de um substantivo ou frase nominal (por exemplo, “Bruno Dantas”) quando um pronome (“ele”) seria suficiente. *

Marcar apenas uma oval.

- Muito ruim
- 1
- 2
- 3
- 4
- 5
- Muito bom

34. **Clareza Referencial**: Deve ser fácil identificar a quem ou a que os pronomes e frases nominais no resumo estão se referindo. Se uma pessoa ou outra entidade for mencionada, deve ficar claro qual é o seu papel na história. Portanto, uma referência não seria clara se uma entidade fosse referenciada, mas sua identidade ou relação com a história permanecesse incerta. *

Marcar apenas uma oval.

Muito ruim

1
2
3
4
5

Muito bom

35. **Foco**: O resumo deve ter um foco; as frases devem conter apenas informações relacionadas ao restante do resumo (sem conteúdo desconexo com o restante do resumo) *

Marcar apenas uma oval.

Muito ruim

1
2
3
4
5

Muito bom

36. **Estrutura e Coerência**: O resumo deve ser bem estruturado e bem organizado. O resumo não deve ser apenas um amontoado de informações relacionadas, mas deve ser construído de frase em frase para um conteúdo coerente de informações sobre o tópico. *

Marcar apenas uma oval.

Muito ruim
1
2
3
4
5
Muito bom

37. Houve algum ato ou fato jurídico gerado no resumo que contradissesse ou extrapolasse o que estava no texto? Caso afirmativo, por favor, descreva abaixo onde ocorreram e qual ponto de contradição e/ou extrapolação que a instrução não dá suporte ao que foi gerado no resumo. *

Resumo da Instrução 3 usando **Método de Modelo Generativo Priorizado por Relevância**:
O texto apresenta um relatório de inspeção do Tribunal de Contas da União sobre a aplicação web TrateCov, desenvolvida e disponibilizada pelo Ministério da Saúde durante o mês de janeiro de 2021. A fiscalização foi solicitada pela Comissão Parlamentar de Inquérito da Pandemia de Covid-19 (CPI-PANDEMIA) do Senado Federal. O relatório conclui que não foram identificados indícios de violação do código-fonte do TrateCov, mas que existem diferenças entre a versão recebida do MS e a versão preservada no Internet Archive. O relatório detalha o funcionamento do TrateCov e os medicamentos recomendados. O texto também apresenta uma matriz de planejamento utilizada na avaliação da aplicação e indica os procedimentos adotados pela equipe de fiscalização.

38. **Gramaticalidade**: O resumo não deve ter datas, formatação interna do sistema, erros de capitalização ou sentenças obviamente agramaticais (por exemplo, fragmentos, componentes ausentes) que dificultem a leitura do texto. *

Marcar apenas uma oval.

Muito ruim
1
2
3
4
5
Muito bom

39. **Não redundância**: Não deve haver repetições desnecessárias no resumo. A repetição desnecessária pode assumir a forma de frases inteiras que se repetem, ou fatos repetidos, ou o uso repetido de um substantivo ou frase nominal (por exemplo, “Bruno Dantas”) quando um pronome (“ele”) seria suficiente. *

Marcar apenas uma oval.

	Muito ruim
1	
2	
3	
4	
5	
	Muito bom

40. **Clareza Referencial**: Deve ser fácil identificar a quem ou a que os pronomes e frases nominais no resumo estão se referindo. Se uma pessoa ou outra entidade for mencionada, deve ficar claro qual é o seu papel na história. Portanto, uma referência não seria clara se uma entidade fosse referenciada, mas sua identidade ou relação com a história permanecesse incerta. *

Marcar apenas uma oval.

	Muito ruim
1	
2	
3	
4	
5	
	Muito bom

41. **Foco**: O resumo deve ter um foco; as frases devem conter apenas informações relacionadas ao restante do resumo (sem conteúdo desconexo com o restante do resumo) *

Marcar apenas uma oval.

Muito ruim
1
2
3
4
5
Muito bom

42. **Estrutura e Coerência**: O resumo deve ser bem estruturado e bem organizado. O resumo não deve ser apenas um amontoado de informações relacionadas, mas deve ser construído de frase em frase para um conteúdo coerente de informações sobre o tópico. *

Marcar apenas uma oval.

Muito ruim
1
2
3
4
5
Muito bom

43. Houve algum ato ou fato jurídico gerado no resumo que contradissesse ou extrapolasse o que estava no texto? Caso afirmativo, por favor, descreva abaixo onde ocorreram e qual ponto de contradição e/ou extrapolação que a instrução não dá suporte ao que foi gerado no resumo. *

44. Quanto tempo total você levou para ler toda instrução e avaliar cada opção de resumo gerada? *

45. Qual a melhor opção de resumo gerado? (Você pode utilizar o botão voltar do formulário para revê-los) *

Marcar apenas uma oval.

- Método de Seleção de Sentenças Mais Relevantes
- Método de Modelo Generativo Priorizado por Relevância

46. (Opcional) Caso haja alguma falha no melhor resumo apontado por você, por favor, detalhe o que faltou nele ou o que ele colocou a mais de forma indevida.

Anexo I

Instrução Processual de exemplo do TCU

TC 014.650/2021-5

Tipo: Representação (com pedido de medida cautelar)

Unidade jurisdicionada: Fundação Universidade do Amazonas

Representante: Esac Engenharia (CNPJ 00.892.637/0001-30)

Representado: Fundação Universidade do Amazonas

Advogado ou Procurador: não há;

Proposta: mérito

INTRODUÇÃO

1. Cuidam os autos de denúncia, com pedido de cautelar *inaudita altera parte*, a respeito de possíveis irregularidades ocorridas na Fundação Universidade do Amazonas (FUAM), relacionadas ao certame licitatório RDC 1/2021, regido pela Lei 12.462, de 4 de agosto de 2011 (lei do Regime Diferenciado de Contratação) com vistas à contratação de remanescente de obra do Bloco 3 pertencente ao Instituto de Educação, Agricultura e Meio Ambiente no campus de Humaitá da Universidade Federal do Amazonas (UFAM), no valor estimado de R$ 2.162.769,52.

BREVE HISTÓRICO

2. Sobre a contratação em comento, em síntese, a representante alega que ela ensejará superfaturamento por parte da futura contratada, uma vez que alguns serviços contemplados na nova licitação já foram executados e que a universidade já tentara relicitar o objeto em questão no RDC 3/2018, com vícios semelhantes, o qual foi revogado pela UFAM em decorrência de representação por ela interposta a este Tribunal (TC 027.595/2018-8).

3. Como visto na instrução pretérita, apesar de ter sido autuada como denúncia, trata-se, na verdade de representação da empresa Esac Engenharia (CNPJ 00.892.637/0001-30), conforme demonstra a peça 41. O tipo do processo já foi inclusive reconhecido pelo Ministro relator, conforme atesta o despacho proferido à peça 51, devendo ser tratado como tal, não havendo razões para manter em sigilo a identidade da representante.

4. Também importa ressaltar que, no mesmo despacho, a representação foi conhecida pelo relator, que acompanhou a proposta da unidade técnica. Já em relação ao pedido de medida cautelar, não obstante tenha concordado com o exame da unidade instrutora e com as medidas saneadoras propostas, divergiu da proposta de adoção imediata da medida cautelar e determinou a realização de oitiva prévia da Fundação Universidade do Amazonas, para que se pronunciasse em cinco dias sobre:

> 14.2.1. elevado valor estimado para a contratação (R$ 2,16 milhões), se comparado com os aproximados R$ 760 mil (base abril/2014, ou R$ 1,11 milhão se atualizados para dezembro/2020), correspondentes a 13,4% do total original, apontados como necessários para a conclusão das obras no parecer contido na peça 32 destes autos;
>
> 14.2.2. inclusão de itens que já foram executados, a exemplo de: instalação de vidros, construção de reservatório elevado e instalação elétrica;
>
> 14.2.3. demais informações que julgar necessárias;

5. Além disso, os demais itens da oitiva solicitavam uma série de justificativas técnicas, acompanhadas dos respectivos elementos comprobatórios, a exemplo dos quantitativos de serviços realizados aproveitáveis e não aproveitáveis executados pela Esac engenharia, empresa representante. Também foram solicitadas fotografias, memórias de cálculo e outros elementos de convicção que permitam concluir pela necessidade de o item integrar o orçamento.

6. A universidade tomou ciência da decisão (peças 57 a 59) e enviou suas justificativas com documentos acostados entre as peças 60 e 64, sendo que na peça 63 consta um item não digitalizável, uma planilha em formato ".xlsx", em resposta à sugestão da unidade técnica de preenchimento da planilha da peça 46, a qual foi acolhida pelo Ministro relator (peça 51, p. 2-3). Essa planilha contém uma lista dos serviços mais importantes do RDC 1/2021, com uma série de informações que justifiquem a sua inclusão no certame. As respostas da universidade serão analisadas a seguir.

EXAME TÉCNICO

7. O reitor da universidade, inicia a sua argumentação (peça 60) informando que o laudo apresentado pela representante apresenta divergências de informações, se comparados os percentuais levantados por ela ao estado real da obra. Informa que encaminhou em anexo documentos que comprovam essa condição e os problemas ocorridos ao longo do contrato.

8. Destaca que o castelo d'água deverá ser demolido e reconstruído, devido ao desaprumo observado, perceptível a olho nu. Acrescenta que foi realizado levantamento topográfico para a confirmação do desaprumo e os resultados foram apresentados a engenheiro do departamento de engenharia, que informou haver necessidade de demolição, pelo risco de tombamento do reservatório, se for preenchido. Também indicou os anexos com os referidos pareceres.

9. Outro item reportado é a instalação de elevador, que não atendeu às especificações técnicas de projeto e que não foi aceito pela última fiscalização. Por isso, os valores referentes à instalação foram incluídos entre as verbas que a construtora deverá restituir aos cofres da União.

10. O reitor também afirma que faltam equipamentos na subestação, como disjuntores de alta tensão, quadros, barramentos, gabinete e demais elementos necessários ao funcionamento do equipamento.

11. Acrescenta que está em andamento o processo de apuração de responsabilidades pelos problemas ocorridos, sendo que o TCU está acompanhando esses procedimentos, por meio do TC 001.267/2019-1, de responsabilidade da Secex Educação. Informa que no processo de sindicância já foi evidenciado que os valores pagos à construtora não condizem com o que foi efetivamente executado.

12. Ressalta a existência de diversas patologias na edificação como infiltrações, infestação por cupins e mofo, que têm causado a perda de muitos serviços executados, além da ocorrência de furtos de elementos já concluídos da obra. Quanto a isso, anexou à documentação o boletim de ocorrência do fato.

13. Na peça seguinte (peça 61), o reitor informa o encaminhamento da planilha contendo as informações solicitadas pelo item 14.3.1 do despacho ministerial. Além dos argumentos já desenvolvidos no ofício anterior, ressalta neste as diferenças de valores em relação a data base dos orçamentos (de 2014 para 2021). Complementa que outro fator a ser considerado é o desconto oferecido pela ora representante, de 13,68%, desconto esse que não é contemplado em orçamento referencial baseado em sistemas oficiais.

14. Na peça 64, o reitor retoma uma série de argumentos já enfrentados pela instrução anterior. De novidade, há a resposta ao questionamento proferido no item 14.4.1.2 do despacho do Ministro relator, onde é solicitada a manifestação quanto aos possíveis impactos de determinação do TCU no sentido de anular o RDC 1/2021.

15. Nessa resposta, o reitor colaciona texto redigido pela Diretora do Instituto de Educação, Agricultura e Ambiente – IEAA, Professora Ana Cláudia Fernandes Nogueira, uma vez que aquele instituto seria o mais impactado pela demora na conclusão das obras. Em resumo, a professora narra toda a história do instituto e descreve as dificuldades de atuação em edifício doado pela Prefeitura de Humaitá, que submete professores, alunos e funcionários a situações improvisadas, prejudicando o ensino e a pesquisa de maneira dramática, frisando que a não entrega do bloco 3, obra objeto deste

processo, continuará afetando os trabalhos administrativos e acadêmicos, bem como a qualidade do ensino prestado.

16. Por fim, o reitor acrescenta também o depoimento do Diretor do Departamento de Engenharia, quanto à eventual anulação do RDC 1/2021:

> Os serviços referentes a construção do bloco 03 do IEAA, foram paralisados em outubro de 2017, nesse período o bloco passou a ter problemas, naturais, de desgaste e depreciação pela falta de manutenção, além de ter passado por eventos de furto, depredação e infestação de cupins. A falta prolongada de uso do prédio causa danos significativos na sua estrutura física, tais como, infiltrações, oxidações de metais, danos em materiais de divisória e MDF. O cancelamento do processo licitatório e a paralização dos serviços, causará mais danos ao prédio, pela ação do tempo, acarretando mais custos físicos e estudantil para comunidade acadêmica do Município de Humaitá.

Análise

17. A peça 60 contém todos os relatórios mencionados pelo reitor. O relatório de vistoria de 2017 (peça 60, p. 5-23) já indicava um descompasso entre as planilhas de medição, fornecidas pela construtora, ora representante, e a execução evidenciada no local. Já naquela ocasião havia sido reportado o pagamento de serviços não executados. Segundo o relatório, o percentual medido pela contratada era de 86,41% e o efetivamente executado era de 65,39%. Conforme bem observado pela instrução precedente (peça 48, item 34) o parecer contratado pela Esac (peça 32) deve ser admitido com cautela.

18. Na verdade, com os novos elementos trazidos aos autos, já é possível afirmar que o parecer realizado pela empresa Agathon não corresponde à versão atual da obra e não contempla a qualidade dos serviços realizados.

19. O mesmo relatório de vistoria 1/2017 também já reportava, além de diversos problemas construtivos, a ausência de responsável técnico ou engenheiro residente durante o período da vistoria, bem como da ausência do diário de obras.

20. A ausência de engenheiro residente já havia sido reportada em notificação emitida em 2016 (peça 60, p. 77-78). Nesta notificação, além do atraso significativo, são reportados o atraso da folha de pagamento, e o número insuficiente de funcionários na obra, entre outras irregularidades. A peça contém ainda outras notificações com reiterações e manifestações semelhantes.

21. Outro relatório relevante é o de número 1/2020, sobre vistoria realizada em setembro de 2020 (peça 60, p. 184-223). Nele é constatada a intensa deterioração de diversos serviços, como elevador exposto a intempéries; inúmeros ninhos de cupins, espalhados por todo o bloco; existência de fungos e umidade excessiva em divisórias; além do furto de diversos elementos como bancadas de granito, bacias sanitárias, folhas de portas, entre outros elementos. Como sugestão dos engenheiros vistoriadores estão a permanência de vigilância no local e a dedetização do bloco.

22. As fotografias anexas ao relatório, e também copiadas na peça 62, confirmam a versão do reitor e do relatório, e demonstram uma edificação bastante danificada pelo abandono, pelas intempéries e pelo vandalismo.

23. Outro documento importante é o parecer técnico indicando o desaprumo do reservatório d'água, indicando uma execução bastante deficiente (peça 60, páginas 224 a 229). Na sequência, foi anexado o parecer indicando a demolição do reservatório. Os pareceres seguintes versam sobre os problemas do elevador e as instalações elétricas, todos indicando a necessidade de complementações e substituições. Por fim, a peça contém um boletim de ocorrência autuado junto à polícia civil do Amazonas (p. 248-249).

24. Todos esses pareceres confirmam os relatos apresentados pela UFAM e indicam a real necessidade de substituição do castelo d'água, do elevador e a reposição de peças da subestação e demais insumos para as instalações elétricas.

25. Essas informações inviabilizam o parecer contratado pela Esac à peça 32. Nesse parecer, é pressuposto que a integralidade dos serviços executados pela Esac pode ser aproveitada. Um bom exemplo é que o parecer nada comenta sobre o reservatório, apenas informando que está 100% concluído e que teve valor contratado de R$ 227.173,66 (peça 32, p. 12) com o desconto oferecido pela empresa na licitação e mencionado na resposta do reitor. O mesmo ocorre com diversos outros serviços que necessitam ser refeitos, a exemplos das portas destruídas pelos cupins e dos forros com infiltração.

26. A planilha anexada à peça 63 foi transformada em arquivo com extensão ".pdf" e juntada à peça 67, de modo que fique facilmente visível a quem tem acesso ao processo. Compulsando essa planilha com a sugestão encartada à peça 46, observa-se que todos os elementos estão presentes, justificando cada um dos itens sobre o motivo da sua inclusão na planilha do RDC 1/2021: se é necessário reparo, em razão de deterioração, ou reposição por furto, ou demolição por defeitos construtivos, entre outras possibilidades. Em outras palavras, entende-se que a referida planilha atende ao solicitado pelo Ministro relator em seu item 14.3.1.

27. Essa planilha, em conjunto com os relatórios anexados aos autos, especialmente aquele elaborado em setembro de 2020, assim como as fotografias, permitem concluir pelo atendimento às demais solicitações encaminhadas pelo despacho da peça 51. Considerando os serviços que efetivamente precisam ser executados, dadas as condições atuais da edificação e a atualização dos referenciais de preços, pode-se afirmar que os itens financeiramente mais relevantes da planilha orçamentária, presentes na peça 67, estão de acordo com os preços de mercado. Releva informar que é esperado ainda algum desconto por parte das empresas interessadas em concluir a obra.

28. Sobre os depoimentos dos professores da universidade, são argumentos que deveriam sensibilizar todos os responsáveis pela condução de obras públicas no país. É compreensível que a falta de estrutura adequada para as atividades de ensino e pesquisa prejudiquem os usuários da universidade e, em última instância, a população que poderia ser beneficiada indiretamente pelos trabalhos ali realizados e pelos profissionais ali formados.

29. Resta saber os motivos e os responsáveis que levaram a essa situação apresentada nos presentes autos. Os problemas com essa edificação já são antigos e certamente contou com a contribuição de muitos agentes. Considerando que as ações tomadas pela Universidade para o ressarcimento dos prejuízos já verificados estão sendo acompanhadas pela Secex Educação, no âmbito do TC 001.267/2019-1, reputa-se necessário informar aquela unidade que, quanto à retomada da obra do bloco 3 em si, a UFAM conseguiu demonstrar a adequação de sua planilha orçamentária à situação real da edificação, podendo-se apensar este processo àquele.

30. Assim, resta comprovada a ausência dos requisitos do *periculum in mora* e do *fumus boni iuris*, que indiquem a paralisação cautelar do RDC 1/2021. Ademais, as justificativas apresentadas pela universidade são suficientes para indicar o prosseguimento do certame, sem prejuízo de a Secex Educação continuar acompanhando as medidas tomadas pela instituição para o ressarcimento dos prejuízos apurados e a responsabilização dos agentes envolvidos com as irregularidades.

CONCLUSÃO

31. Trata-se de análise de oitiva prévia e diligência da Fundação Universidade do Amazonas (FUAM), sobre supostas irregularidades no RDC 1/2021, que tem como objeto a contratação da execução do remanescente de obras do bloco 3, no campus de Humaitá. Não foi verificada a existência dos pressupostos que indicassem a adoção de medida cautelar.

32. A documentação acostada aos autos e as justificativas apresentadas foram suficientes

para comprovar a adequação da planilha de referência do certame, pelo menos em relação aos itens financeiramente mais relevantes (peça 67). Muitos dos serviços indicados como realizados pela representante ou foram perdidos por deterioração, ou foram malfeitos (a exemplo do castelo d'água) ou ainda roubados ou vandalizados. Assim, no mérito, a representação pode ser considerada improcedente, pois o alegado superfaturamento que beneficiaria a futura contratada não foi comprovado.

33. Assim, entende-se recomendável a continuidade do RDC 1/2021, sem prejuízo de a Secex Educação continuar acompanhando as medidas tomadas pela instituição para o ressarcimento dos prejuízos apurados e a responsabilização dos agentes envolvidos com as irregularidades ocorridas nos contratos anteriores, no âmbito do TC 001.267/2019-1.

PROPOSTA DE ENCAMINHAMENTO

34. Ante todo o exposto, submetem-se os autos à consideração superior, propondo:

a) **alterar** o tipo processual de denúncia para representação;

b) **conhecer** da **representação**, satisfeitos os requisitos de admissibilidade constantes no art. 113, § 1º, da Lei 8.666/1993, c/c os arts. 235 e 237, VII, do Regimento Interno deste Tribunal, e no art. 103, § 1º, da Resolução - TCU 259/2014 para, no mérito, considerá-la improcedente;

c) **indeferir** o pedido de medida cautelar formulado pela representante, pela ausência dos requisitos para tal medida;

d) **apensar definitivamente**, com fulcro no arts. 36 e 37 da Resolução - TCU 259/2014, alterada pela Resolução-TCU 321/2020, o presente processo ao TC 001.267/2019-1, posto que há relação de conexão parcial entre eles e se mostra conveniente a tramitação conjunta;

e) **informar** à representante e à Fundação Universidade do Amazonas do acórdão que vier a ser proferido, destacando que o relatório e o voto que fundamentam a deliberação ora encaminhada podem ser acessados por meio do endereço eletrônico www.tcu.gov.br/acordaos;

SeinfraUrbana, 11 de agosto de 2021

(Assinado eletronicamente)

Marcos Donizete Machado

AUFC– Mat. 9435-8