Pontifícia Universidade Católica
do Rio de Janeiro

Dayson Nywton Corrêa Rodrigues do Nascimento

Sumarização de artigos científicos em Português no domínio da Saúde

Dissertação de Mestrado

Dissertação apresentada como requisito parcial para a obtenção do grau de Mestre pelo Programa de Pós–graduação em Informática da PUC-Rio.

Orientador : Prof. Hélio Côrtes Vieira Lopes
Coorientador: Dr. Fernando Alberto Correia dos Santos Junior

Rio de Janeiro
setembro de 2023

Pontifícia Universidade Católica
do Rio de Janeiro

Dayson Nywton Corrêa Rodrigues do Nascimento

Sumarização de artigos científicos em Português no domínio da Saúde

Dissertação apresentada como requisito parcial para a obtenção do grau de Mestre pelo Programa de Pós–graduação em Informática da PUC-Rio. Aprovada pela Comissão Examinadora abaixo:

Prof. Hélio Côrtes Vieira Lopes
Orientador
Departamento de Informática – PUC-Rio

Dr. Fernando Alberto Correia dos Santos Junior
Coorientador
Pontifícia Universidade Católica do Rio de Janeiro – PUC-Rio

Prof. Marcos Kalinowski
Pontifícia Universidade Católica do Rio de Janeiro – PUC-Rio

Prof. Cássio Freitas Pereira de Almeida
ENCE

Rio de Janeiro, 22 de setembro de 2023

Dayson Nywton Corrêa Rodrigues do Nascimento

Graduou-se em Ciência da Computação pela Universidade Universidade Federal do Maranhão (São Luís). Atualmente mestrando no Departamento de Informática da PUC-Rio.

Ficha Catalográfica

Nascimento, Dayson

Sumarização de artigos científicos em Português no domínio da Saúde / Dayson Nywton Corrêa Rodrigues do Nascimento; orientador: Hélio Côrtes Vieira Lopes; coorientador: Fernando Alberto Correia dos Santos Junior. – 2023.

72 f: il. color. ; 30 cm

Dissertação (mestrado) - Pontifícia Universidade Católica do Rio de Janeiro, Departamento de Informática, 2023.

Inclui bibliografia

1. keywordpre – Teses. 2. keywordpre – Teses. 3. Sumarização abstrativa. 4. Large Language Models. 5. Artigos científicos. 6. Português. 7. Fine-tuning. I. Côrtes Vieira Lopes, Hélio. II. Correia, Fernando. III. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Informática. IV. Título.

CDD: 004

À minha família.

Agradecimentos

Às vezes sinto como se eu fosse muito sortudo. Como se tudo desse certo. Mesmo quando parece errado, é como se houvesse uma conspiração do bem que me levasse a cometer erros por conta de algum aprendizado novo, a conhecer as pessoas certas no momento certo (mesmo que eu não saiba que seja o certo). Parece ser algo divino que me trouxe até aqui. E acredito que seja! Por isso, quero agradecer primeiro a Deus pela vida, pelo erros, pelos acertos, por estar aqui, por realizar mais um sonho, pela minha família, pelos meus amigos, por esse trabalho, por ser privilegiado! Me sinto tão grato que eu escreveria 10 páginas de agradecimento...

Ao meu orientador, Hélio Lopes, pela oportunidade, confiança e apoio durante o desenvolvimento dessa pesquisa. Ao meu amigo e co-orientador Fernando Correia pela ajuda e orientação especialmente nessa etapa final.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001. À PUC-Rio, pela bolsa de isenção das anuidades e pelo acesso à uma infraestrutura de pesquisa e a um enriquecedor ambiente acadêmico.

Aos meus pais pelos ensinamentos, por me tornarem quem sou, por seus sacrifícios para que eu tivesse educação de qualidade e pudesse ter o privilégio que eles não tiveram. Aos meus irmãos que sempre foram meus parceiros de sangue e de vida. À minha esposa, Caroline Nascimento, que topou toda a nossa mudança de vida para que eu alcançasse meu sonho que é esse Mestrado na PUC-Rio.

Aos meus amigos, em especial a Dalai, que me relembrou desse sonho, e aos do mestrado, Vinicius e Daniel, que sempre me mantiveram próximo e motivado para acabar.

Resumo

Nascimento, Dayson; Côrtes Vieira Lopes, Hélio; Correia, Fernando. **Sumarização de artigos científicos em Português no domínio da Saúde**. Rio de Janeiro, 2023. 72p. Dissertação de Mestrado – Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

Neste trabalho, apresentamos um estudo sobre o fine-tuning de um LLM (Modelo de Linguagem Amplo ou Large Language Model) pré-treinado para a sumarização abstrativa de textos longos em português. Para isso, construímos um corpus contendo uma coleção de 7.450 artigos científicos na área de Ciências da Saúde em português. Utilizamos esse corpus para o *fine-tuning* do modelo BERT pré-treinado para o português brasileiro (BERTimbau). Em condições semelhantes, também treinamos um segundo modelo baseado em Memória de Longo Prazo e Recorrência (LSTM) do zero, para fins de comparação. Nossa avaliação mostrou que o modelo ajustado obteve pontuações ROUGE mais altas, superando o modelo baseado em LSTM em 30 pontos no F1-score. O *fine-tuning* do modelo pré-treinado também se destaca em uma avaliação qualitativa feita por avaliadores a ponto de gerar a percepção de que os resumos gerados poderiam ter sido criados por humanos em uma coleção de documentos específicos do domínio das Ciências da Saúde.

Palavras-chave

Sumarização abstrativa; Large Language Models; Artigos científicos; Português; Fine-tuning.

Abstract

Nascimento, Dayson; Côrtes Vieira Lopes, Hélio (Advisor); Correia, Fernando (Co-Advisor). **Sumarization of health science papers in Portuguese**. Rio de Janeiro, 2023. 72p. Dissertação de Mestrado – Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

In this work, we present a study on the fine-tuning of a pre-trained Large Language Model for abstractive summarization of long texts in Portuguese. To do so, we built a corpus gathering a collection of 7,450 public Health Sciences papers in Portuguese. We fine-tuned a pre-trained BERT model for Brazilian Portuguese (the BERTimbau) with this corpus. In a similar condition, we also trained a second model based on Long Short-Term Memory (LSTM) from scratch for comparison purposes. Our evaluation showed that the fine-tuned model achieved higher ROUGE scores, outperforming the LSTM based by 30 points for F1-score. The fine-tuning of the pre-trained model also stands out in a qualitative evaluation performed by assessors, to the point of generating the perception that the generated summaries could have been created by humans in a specific collection of documents in the Health Sciences domain.

Keywords

Abstractive Summarization; Large Language Models Fine-tuning; Scientific papers; Portuguese; Fine-tunin.

Sumário

1 Introdução **1**
1.1 Contribuições 3
1.2 Organização do trabalho 4

2 Revisão de Literatura **5**
2.1 O que é sumarização 5
2.2 Técnicas e Abordagens de Sumarização Abstrativa 6
2.2.1 O que são Redes Neurais 7
2.2.2 Sumarização de textos 8

3 Trabalhos Relacionados **13**
3.1 Desenvolvimento dos modelos 13
3.2 Domínio e corpus 14
3.3 Métodos de avaliação 16

4 Construção do Corpus **19**
4.1 Coleta de dados 19
4.1.1 Web Crawler 20
4.1.2 Tratamento de dados 21
4.2 Análise do Corpus 22
4.2.1 Descrição dos Resumos 22
4.2.2 Descrição dos Textos Principais 25

5 Geração dos resumos **27**
5.1 LSTM *from the scratch* 27
5.2 *Fine-tuning* uma versão do *BERTimbau* 30
5.3 Estratégias de avaliação 31
5.3.1 Avaliação Quantitativa 31
5.3.2 Avaliação Qualitativa com Avaliadores 32

6 Resultados **36**
6.1 Desempenho dos modelos 36
6.2 Avaliação qualitativa 41
6.2.1 Perfil dos participantes 41
6.2.2 Análise das respostas 43

7 Conclusões **46**

Referências bibliográficas **48**

A Material do Estudo **56**
A.1 Termo de Consentimento Livre e Esclarecido 56
A.2 Questionário de Caracterização do Perfil do Participante 59
A.3 Questionário para Captura de Opiniões sobre os resumos 59

Lista de Figuras

Figura 2.1 Diagrama por Raffel et al. (2020) representando o framework *text-to-text*. "T5"é uma referência ao modelo utilizado baseado em Transformers. 7
Figura 2.2 Representação de uma Rede Neural MLP (*Multilayer Perceptron*) por Raschka (2018). 8
Figura 2.3 Representação de uma Rede Neural Recorrente (RNN) por Mittal (2019). 9
Figura 2.4 Representação de uma Rede Neural de Memória de Longo Prazo (LSTM) por Mittal (2019). 10
Figura 2.5 Representação de um Transformer por Vaswani et al. (2017). 11

Figura 4.1 Metodologia utilizada nesse trabalho. 19
Figura 4.2 Nuvem de palavras com as palavras mais frequentes dos resumos contidos no corpus. 24
Figura 4.3 Histograma do tamanho dos resumos. 24
Figura 4.4 Nuvem de palavras com as palavras mais frequentes dos textos principais dos artigos. 25
Figura 4.5 Histograma tamanho do corpo principal dos artigos. 26

Figura 5.1 Visão geral do modelo de Cohan et al. (2018). Representação do *step* 3 de decodificação. A menção a RNN sempre se refere às LSTMs Bidirecionais. 28
Figura 5.2 Exemplo de cálculo do *Recall* utilizando o Rouge-N (Briggs, 2021). 32
Figura 5.3 Exemplo de cálculo do *Recall* utilizando o Rouge-L (Briggs, 2021). 32
Figura 5.4 Etapas do procedimento para a avaliação dos textos gerados. 33

Figura 6.1 Referência para os resumos gerados da Figura 6.2. 37
Figura 6.2 Exemplo 01 dos resumos gerados pelos dois modelos. 38
Figura 6.3 Referência para os resumos gerados da Figura 6.4. 39
Figura 6.4 Exemplo 02 dos resumos gerados pelos dois modelos. 39
Figura 6.5 Referência para os resumos gerados da Figura 6.6. 40
Figura 6.6 Exemplo 03 dos resumos gerados pelos dois modelos. 41
Figura 6.7 Distribuição dos participantes por grau de escolaridade. 42
Figura 6.8 Média de todas as 36 respostas sobre os resumos elaborados tanto por humanos quanto pela IA em todos os grupos (1, 2, 3 e 4). 43
Figura 6.9 Média das respostas sobre os resumos do grupo 2 (elaborados apenas pela IA). 44
Figura 6.10 Média das respostas sobre a semelhança entre os conteúdos apresentados pelos dois resumos. 45

Figura A.1 Comparação entre Resumos 01 e 02 60
Figura A.2 Avaliação do Primeiro Resumo 60
Figura A.3 Avaliação do Segundo Resumo 60
Figura A.4 Similaridade entre Resumos. As respostas padronizadas contêm 7 opções na Escala de Likert. 60

Lista de Tabelas

Tabela 3.1 Ranking dos 10 unigramas e bigramas mais frequentes nos textos principais do corpus. 18

Tabela 4.1 Análise morfológica das palavras contidas no dataset. 22
Tabela 4.2 Mais detalhes sobre o corpus 23
Tabela 4.3 Ranking dos 10 unigramas e bigramas mais frequentes nos resumos do corpus. 24
Tabela 4.4 Ranking dos 10 unigramas e bigramas mais frequentes nos textos principais do corpus. 25

Tabela 6.1 ROUGE-1, ROUGE-2 e ROUGE-L Recall e F1-Score como resultados no nosso conjunto de teste em português. 37

Lista de Abreviaturas

CNN - *Convolutional Neural Network* (Redes Neurais Convolucionais) IA – Inteligência Artificial

ML – Machine Learning (Aprendizado de Máquina)

MLP - *Multilayer Perceptron*

NLP – *Natural Language Processing* (Processamento de Linguagem Natural)

LLM – *Large Language Model*

RNN – Rede Neural Recorrente

LSTM – Long Short-Term Memory

BERT – Bidirectional Encoder Representations from Transformers

TLDR – Too Long, Didn't Read (*muito longo, não li*)

BLEU – *Bilingual Evaluation Understudy*

Seq2Seq – *Sequence to Sequence* (Sequência para Sequência)

1
Introdução

Atualmente, temos grandes coleções de documentos acessíveis, espalhadas por diferentes domínios e assuntos, de artigos científicos, notícias, à documentos jurídicos. Em contraponto a recuperação de informações relevantes em tais coleções, especialmente aquelas majoritariamente compostas por documentos longos e verbosos, como relatórios técnicos e artigos científicos, ainda representa um desafio. Motores de busca podem ajudar os usuários a filtrar um conjunto de documentos com potencial relevância. No entanto, um refinamento desse resultado implica em uma leitura atenta desses documentos, capturando as informações mais relevantes de cada documento. Nesse contexto, a sumarização de texto vem como uma solução elegante para facilitar esse processo extração de informações relevantes.

A sumarização de texto é uma técnica de Processamento de Linguagem Natural (NLP) que visa condensar a ideia central original de um texto, mantendo suas informações mais importantes (Liu e Lapata, 2019). Gupta e Gupta (2019) categoriza as técnicas de sumarização em abordagens extrativas e abstrativas. A abordagem extrativa seleciona as sentenças mais importantes do texto e as organizam de forma coerente em um texto curto. Não há a criação de conteúdo novo, mas o recorte e organização de sentenças já existentes no texto original. Enquanto a abstrativa, gera um resumo utilizando novas palavras e sentenças, podendo de fato gera novas frases (até mesmo variando palavras) para produzir um resumo coerente.

Neste trabalho, utilizaremos a sumarização abstrativa. Os algoritmos para sumarização abstrativa são projetados para analisar profundamente o texto completo e sua semântica, criando uma representação vetorial que gera um texto mais curto. Durante o treinamento, o modelo aprende a partir do texto completo como entrada e seu resumo como alvo, e sua versão final pode gerar resumos coerentes e semelhantes aos escritos por humanos.

Resultados recentes, como os apresentados por Zolotareva et al. (2020); Narayan et al. (2018); Ma et al. (2021), mostram que a utilização de modelos de aprendizado profundo (*deep learning*) tem se mostrado uma abordagem eficaz para várias tarefas de sumarização abstrativa (mais detalhes serão apresentados no Capítulo 2). Alguns desses trabalhos são baseados em modelos pré-

treinados como o *Bidirectional Encoder Representations from Transformers* (BERT) (Devlin et al., 2019), *Chat Generative Pre-trained* (GPT) (Radford et al., 2018) e o BARD (Thoppilan et al., 2022), que são projetados para analisar profundamente grandes coleção de textos (grandes corpus) e aprender a capturar importantes informações relativas a semântica desses textos. No contexto de sumarização, eles têm a capacidade de aprender um padrão "humano" de geração de resumos e repetir esse padrão para a geração de resumos coerentes e próximos aos escritos por humanos.

Devido ao alto custo do processo de treinamento desses modelos, eles são desenvolvidos ou financiados por grandes empresas de tecnologia. Um aspecto a ser visto é o interesse crescente em aplicar a técnica de pequenos ajustes, conhecida como *fine-tuning*, para adaptar essas redes neurais a tarefas ou domínios específicos por um baixo custo de treinamento.

O *fine-tuning* envolve a reconfiguração de um modelo pré-treinado utilizando dados adicionais específicos do domínio ou da tarefa em questão. Isso permite que o modelo seja refinado para realizar tarefas específicas de forma mais precisa e eficiente. As empresas de tecnologia têm sido pioneiras nesse processo de *fine-tuning*, aproveitando suas grandes capacidades computacionais e vastos conjuntos de dados. Essa abordagem tem se mostrado eficaz para aplicações em várias áreas, permitindo que os modelos pré-treinados sejam adaptados para atender a requisitos particulares e fornecer resultados mais relevantes e precisos para diferentes cenários, mesmo com menor uso de hardware e com datasets enxutos, bem menores que os utilizados no pré-treino. A capacidade de *fine-tuning* é uma das razões pelas quais esses *Large Language Models* (LLM), como o ChatGPT [1], têm gerado grande interesse e têm sido utilizados em uma variedade de aplicações no campo do processamento de linguagem natural como o Google Tradutor [2], corretores automáticos de texto, entre outros.

Embora muito progresso tenham sido alcançados com *fine-tuning* em modelos pré-treinados aplicados a textos em inglês (Raffel et al., 2020; An et al., 2021; Zolotareva et al., 2020; Liu e Lapata, 2019; Yasunaga et al., 2019; Xiao e Carenini, 2019; Cohan et al., 2018), idiomas menos comuns têm recebido menos atenção. Este trabalho concentra-se em produzir sumarizações abstrativas de textos longos em português utilizando Redes Neurais. Para isso, em primeiro lugar, construímos um corpus reunindo uma coleção de 7.450 artigos científicos em português no domínio da Saúde Pública. Com esse corpus, realizamos o *fine-tuning* de um modelo pré-treinado BERT para o português brasileiro, o

[1] ChatGPT: `https://chat.openai.com/`

[2] `https://translate.google.com/`

BERTimbau (Souza et al., 2020). Além disso, também treinamos um segundo modelo baseado em Long Short-Term Memory (LSTM) do zero para fins de referencial de comparação.

Ambos os modelos selecionados foram treinados em condições semelhantes e, como esperado, nossas avaliações mostraram que o BERTimbau após o fine-tuning alcançou melhores resultados. Essa avaliação foi feita utilizando o ROUGE, uma métrica automática de comparação de textos, mais informações na Seção 3.3. Este estudo, portanto, lança luz sobre a sumarização abstrativa em português, destacando a importância do pré-treinamento de modelos (como o BERTimbau) em corpora extensos e variados para obter um desempenho aprimorado em tarefas específicas de domínio.

Além dos resultados mencionados, também progredimos na avaliação por meio de um estudo qualitativo feito com a participação de 36 avaliadores. Esses resultados sugerem que o *fine-tuning* de LLMs pode ser uma abordagem promissora para aprimorar a capacidade de resumir textos em português e em outras línguas menos comuns.

A motivação para focar na Saúde Pública neste projeto foi devido ao recente cenário de pandemia. Nesse período, houve um aumento significativo (92%) nas novas publicações de artigos científicos na área (Else, 2020). Em contrapartida, se tornou mais urgente a proposta de ferramenta de inteligência artificial capazes de apoiar os profissionais de saúde.

Neste trabalho, a pesquisa foi conduzida buscando responder à questão de pesquisa (QP) **Como resumir automaticamente documentos técnicos de Ciências da Saúde em português?** Que pode ser quebrada em 3 outras perguntas:

> ***QP1****: Como gerar resumos coerentes e informativos de artigos científicos escritos em português utilizando uma abordagem abstrativa?*
>
> ***QP2****: Como viabilizar estratégias comumente utilizadas no inglês para o português?*
>
> ***QP3****: Como avaliar de forma qualitativa resumos gerados utilizando uma abordagem abstrativa?*

1.1 Contribuições

Nossa principal contribuição com este trabalho é (i) desenvolver e disponibilizar um modelo de sumarização de textos longo em português que possa ser replicado pela comunidade científica, incluindo (ii) um corpus dedicado a

sumarização na área de Ciências da Saúde que foi disponibilizado publicamente e servirá como referência para esse projeto bem como para qualquer outro estudo na área; (iii) uma proposta de processo de treino e teste de modelos para geração de resumos seguindo a abordagem abstrativa para textos em português; e (iv) uma apresentação de proposta de avaliação qualitativa sobre os resumos gerados pelos modelos aqui propostos.

1.2 Organização do trabalho

A estrutura deste trabalho é a seguinte. No Capítulo 2, apresentamos os principais estudos que identificamos em nossa revisão de literatura, relacionando-os a um embasamento teórico. Nosso objetivo é traçar um paralelo de como as redes neurais e as técnicas de sumarização têm progredido juntas ao longo do tempo. No Capítulo 3, descrevemos os trabalhos mais relevantes relacionados a este. O Capítulo 4 detalha a abordagem que adotamos para resolver o desafio da sumarização de textos extensos em português, focando na coleta do nosso corpus em português, o tratamento e uma análise desses dados. No Capítulo 5 tratamos das escolhas dos modelos para a tarefa de sumarização, dos seus treinamentos e dos métodos de avaliação das previsões geradas por esses modelos, tanto de maneira automática quanto manual. No Capítulo 6, compartilhamos os resultados que obtivemos ao longo do processo. Por fim, no Capítulo 7, apresentamos nossas conclusões, considerações finais e delineamos possíveis direções para futuros trabalhos.

2
Revisão de Literatura

No nosso mundo cada vez mais inundado de informações, surge uma pergunta recorrente: é realmente necessário mergulhar em todos os textos que aparecem como resultados nas nossas buscas? Como podemos filtrar de maneira eficaz e inteligente o que realmente queremos, especialmente quando nos deparamos com textos densos em terminologias técnicas, como relatórios ou manuais?

A partir dessa perspectiva, o nosso foco foi direcionado à busca por algoritmos e técnicas capazes de realizar uma tarefa fundamental: a sumarização automática de textos longos, de um domínio específico e escritos em português. Nosso objetivo é gerar resumos que se aproximem da qualidade de resumos produzidos por seres humanos, informativos, com coesão e legibilidade.

Sendo assim, neste capítulo, apresentamos uma revisão da literatura sobre a sumarização de texto, abordando pesquisas relevantes relacionadas ao resumo abstrativo de textos que possam ser aplicados à língua portuguesa. A revisão está organizada em subseções que abrangem diferentes aspectos da sumarização de texto.

2.1
O que é sumarização

Como mencionado no Capítulo 1, existem duas principais abordagens para a sumarização de texto: a sumarização extrativa e a sumarização abstrativa. Por um lado, na sumarização extrativa, trechos literais do texto original são selecionados e rearranjados para formar o resumo. Antes de realizar a seleção, essa técnica avalia a importância estatística de cada trecho do texto, classificando-os por relevância para construir um resumo coerente e que mantenha a ideia central do documento-fonte. A sumarização extrativa tem sido amplamente estudada e é considerada uma área bem estabelecida de pesquisa (Gupta e Gupta, 2019).

Na sumarização abstrativa, o algoritmo gera novas frases e palavras para formar um resumo coeso e conciso. Essa abordagem requer uma análise mais profunda do texto completo e sua semântica, criando uma representação vetorial que resulta em um texto mais curto e abstrato. Durante a fase de treinamento, o modelo aprende a partir de dois textos: o texto original como

entrada e o resumo humano como alvo. Isso permite que o modelo produza resumos que podem conter vocabulário ligeiramente diferente daquele presente no texto original. A sumarização abstrativa é uma área de pesquisa desafiadora e em constante evolução (Liu e Lapata, 2019).

Gupta e Gupta (2019) realizam uma análise de modelos de sumarização abstrativa que atingiram o estado-da-arte. Embora essa publicação seja de 2019 e não faça referência a resultados recentes, como o surgimento do BERT, ela mapeia várias pesquisas, especialmente aquelas baseadas em modelos RNN e LSTM, categorizando-as de acordo com conceitos encontrados na literatura.

Além disso, a geração de resumos pode ser feita de duas maneiras, dependendo da quantidade de documentos de origem utilizados. No caso da sumarização de documento único, o objetivo é criar um resumo conciso a partir de apenas uma fonte de texto. Já na sumarização de documentos múltiplos, diferentes documentos, mas que abordam tópicos semelhantes, são utilizados como entrada para produzir um resumo que reúne as ideias principais de todos eles em um espaço reduzido. Essa última abordagem é especialmente útil quando lidamos com conjuntos de documentos relacionados a um mesmo tema (Ibrahim Altmami e El Bachir Menai, 2022).

Adicionalmente, Ibrahim Altmami e El Bachir Menai (2022) conduzem uma análise abrangente sobre a sumarização de artigos científicos. Eles revisam diferentes tipos de abordagens, abrangendo sumarização abstrativa, extrativa e híbrida, demonstrando métodos amplamente empregados na geração automática de resumos.

2.2 Técnicas e Abordagens de Sumarização Abstrativa

Existem diferentes abordagens na literatura que evoluíram ao longo dos últimos anos para resolver os mais diversos problemas de NLP. Para lidar com isso no geral, muitos pesquisadores têm optado por utilizar modelos de aprendizado de máquina, geralmente com redes neurais, para compreender o texto e criar conhecimento para realizar diversas tarefas, como tradução, resposta a perguntas, preenchimento de lacunas de texto ou sumarização. A Figura 2.1 ilustra uma representação melhorada de uma estrutura de *text-to-text*. Essa abordagem conceitualiza o problema como uma tarefa de sequência para sequência (Seq2Seq), em que uma sequência de palavras é mapeada para gerar outra sequência de palavras como resultado (Sutskever et al., 2014).

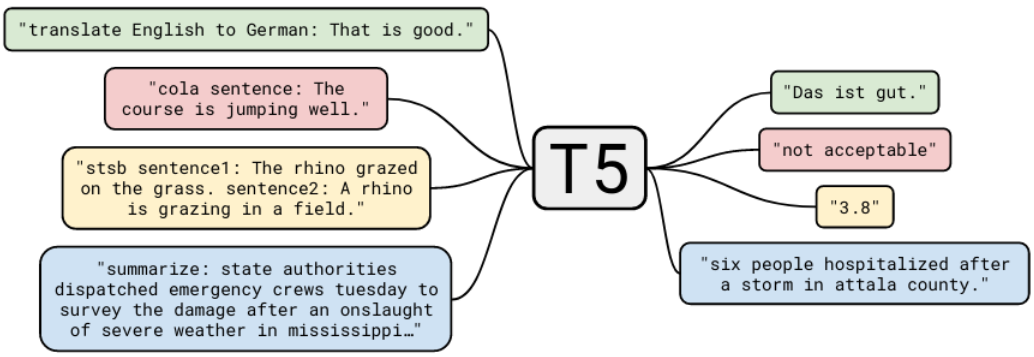


Figura 2.1: Diagrama por Raffel et al. (2020) representando o framework *text-to-text*. "T5"é uma referência ao modelo utilizado baseado em Transformers.

2.2.1 O que são Redes Neurais

O conceito das redes neurais teve sua origem na contribuição de Rosenblatt (1958), que propôs um modelo matemático destinado a imitar as conexões cerebrais presentes em sistemas nervosos. Dentro desse modelo, um elemento chave para os dias de hoje foi introduzido: o neurônio artificial, nomeado *perceptron*, com a finalidade de atuar como um transmissor de informações ao longo de conexões interligadas, espelhando o que acontece biologicamente.

Nessa abordagem, uma entrada é introduzida no sistema e transmitida ao "neurônio"para processamento, resultando em uma saída. O desenho proposto do *perceptron* consiste em três partes essenciais: uma camada de entrada, uma camada ponderada e uma camada de saída.

Esse modelo de neurônio deu origem ao *Multilayer Perceptron* (MLP), uma forma elementar de rede neural. Assim, diversas camadas de neurônios são organizadas em um arranjo interconectado, ilustrado na Figura 2.2. Seu processo de treinamento passa pela otimização dos pesos de tal modo que um conjunto de entradas produza as saídas desejadas. Essa arquitetura habilita, então, a rede para o processamento e transmissão de informações, permitindo a execução de diversas tarefas. Assim, à medida que os métodos de treinamento e o poder computacional avançaram, essa estrutura passou a ter a conexão entre milhões de camadas, evoluindo para o *deep learning*, uma revolução na inteligência artificial (Noriega, 2005).

Com o decorrer do tempo, as Redes Neurais se diversificaram para acomodar uma variedade de tarefas. Novos paradigmas surgiram, incluindo o uso de redes neurais convolucionais (CNNs) (Krizhevsky et al., 2012; Simonyan e Zisserman, 2014) para processamento de imagens e áudio, a adoção de Transformers (Devlin et al., 2019; Vaswani et al., 2017) para processamento de texto e linguagem natural, bem como modelos de múltiplas tarefas, com

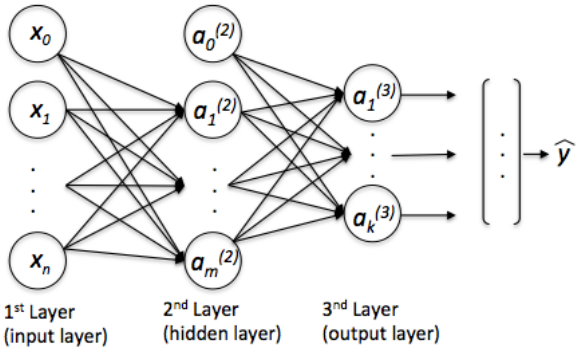


Figura 2.2: Representação de uma Rede Neural MLP (*Multilayer Perceptron*) por Raschka (2018).

a capacidade de extrair informações textuais de vídeos (Vinyals et al., 2015; Kim et al., 2016).

2.2.2 Sumarização de textos

Ao explorarmos a literatura sobre modelos de sumarização, identificamos três principais tipos de Redes Neurais que alcançaram o estado da arte nessa área nos últimos anos. Esses grupos são: Redes Neurais Recorrentes (RNN) (Xiao e Carenini, 2019; Hermann et al., 2015), Redes Neurais de Memória de Longo Prazo (LSTM) Cohan et al. (2018); Yasunaga et al. (2019); An et al. (2021), e o *fine-tuning* de LLMs baseados em Transformers, como é o caso do modelo BERT (Liu e Lapata, 2019; Zolotareva et al., 2020). A história da evolução das redes neurais está diretamente conectada aos modelos que atingiram o estado da arte.

Nesse contexto, as RNNs foram as primeiros a surgir com o objetivo inicial de resumir textos. Essas redes, ilustradas na Figura 2.3, operam da seguinte forma: inicialmente, recebem uma palavra de entrada x_0, gerando uma saída h_0. No próximo passo, ao receber uma nova palavra x_1 junto com a saída anterior h_0, produzem uma nova saída, carregando a informação anterior. Esse processo se repete até que todas as palavras do texto sejam processadas. A partir desse modelo, os codificadores leem o texto palavra por palavra, codificando-o em uma representação espacial. Essa representação preserva o significado capturado no texto original e serve como entrada para um segundo modelo, o decodificador (também usando RNNs), estabelecendo uma conexão entre o codificador e o decodificador. O decodificador é responsável por gerar as representações das novas palavras, que, combinadas, formam o resumo do texto (Hermann et al., 2015).

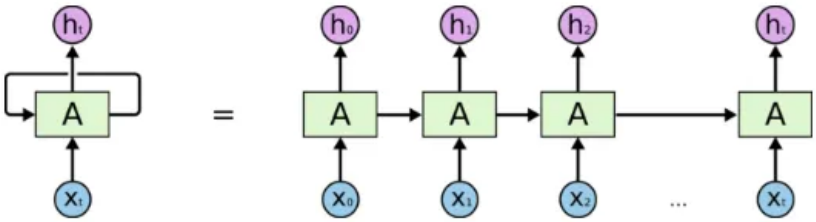


Figura 2.3: Representação de uma Rede Neural Recorrente (RNN) por Mittal (2019).

Em relação aos modelos baseados em RNN, uma vez que esses modelos têm suas limitações em capturar efetivamente dependências de longo prazo em textos extensos, estudos anteriores em sumarização se concentram em documentos curtos, como resumos de artigos de notícias com aproximadamente 1.000 palavras (Hermann et al., 2015).

Apesar disso, um estudo realizado por Xiao e Carenini (2019) propõe uma abordagem de pipeline para gerar resumos extrativos utilizando modelos RNN a partir de documentos longos. Esse método coleta contexto global e local ao longo do texto para comprimir o conteúdo, sem perder informações importantes. Ao combinar codificadores de texto e frases, juntamente com classificadores de frases, o modelo RNN foi capaz de gerar resumos que melhor representam artigos científicos.

Há uma limitação nas RNNs em preservar dependências de longo prazo, visto que elas se baseiam nas palavras imediatamente vizinhas para construir contexto. Restava o desafio de fazer que o contexto das primeiras palavras de um texto longo se mantivesse até a última palavra. Para resolver esse desafio, surgiram as LSTMs. Estas se destacam por sua arquitetura projetada para lidar de maneira mais eficaz com dependências temporais que se estendem por períodos mais extensos (Zolotareva et al., 2020), representadas na Figura 2.4. Seu desenho incorporou *gates*: o de entrada que alimenta a memória da rede, o de esquecimento que descarta aquilo que não deve ser preservado e o de saída que cruza toda essas informações e gera uma saída.

Essa representação trouxe eficiência para as LSTMs em capturar dependências temporais mais amplas e permitiu uma melhor absorção das informações contidas em textos extensos. Isso se tornou um pilar para diversos modelos de sumarização que buscaram alcançar o estado da arte no campo da sumarização (Melis et al., 2017; Cheng e Miyao, 2017; Wang et al., 2018; Chiu e Nichols, 2016). A preferência pela arquitetura LSTM destaca sua eficácia em enfrentar as complexidades das dependências temporais prolongadas, resultando em melhorias substanciais no desempenho da geração automática

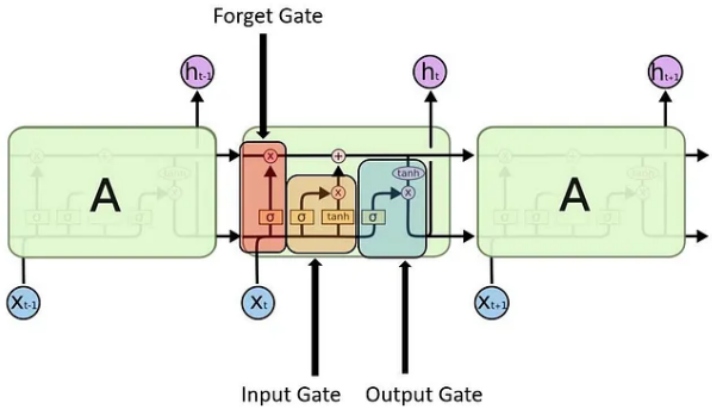


Figura 2.4: Representação de uma Rede Neural de Memória de Longo Prazo (LSTM) por Mittal (2019).

de resumos quando comparadas às RNNs.

Mais recentemente, o Transformer foi criado por Vaswani et al. (2017) como nova arquitetura de redes neurais com o propósito de melhorar pontos de dificuldades das LSTMs como: necessidade de longos tempos de treinamento e pouco paralelismo. Sua arquitetura é mais complexa que as demais redes. A Figura 2.5 mostra a arquitetura do Transformer e seu funcionamento. A entrada é convertida em vetores de palavras (embeddings) e passa por um processo de codificação posicional para capturar a ordem das palavras. Os vetores são então processados pelos blocos de codificação, começando com a camada de atenção *multi-head*, que permite considerar informações de diferentes palavras simultaneamente. Há também uma camada de *Add and normalize* para tratar conexões residuais. Após a atenção *multi-head*, os dados são direcionados para uma camada de *feed forward* e, em seguida, para o decodificador.

No decodificador, o processo é semelhante, mas inclui uma camada de *Masked Multi-Head Attention*, uma versão modificada da autoatenção. Esta máscara impede que a rede acesse informações que não deveria ter, como palavras futuras ao prever uma sequência. Isso garante que a rede não possa ver a palavra-alvo durante a previsão.

Esse novo modelo não trata as palavras em ordem sequencial (apenas guarda informações sobre as posições) e baseia-se em mecanismos de atenção, que dão mais pesos a partes mais relevantes das entradas e menos pesos para outras menos relevantes, e mostrou-se capaz de lidar com longas sequências e capturar relacionamentos de palavras distantes em um texto.

No contexto da arquitetura Transformer, esses mecanismos de atenção são usados para processar as entradas de uma maneira paralela e altamente

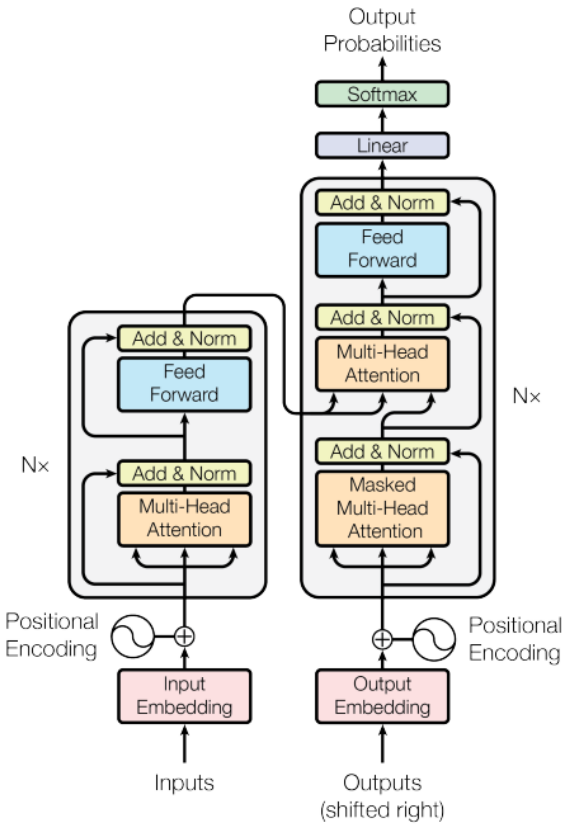


Figura 2.5: Representação de um Transformer por Vaswani et al. (2017).

eficiente, em contraste com abordagens anteriores que dependiam fortemente de operações sequenciais, como recorrências (RNN e LSTM) e convoluções (CNN). Isso contribui para a capacidade do Transformer de lidar melhor com sequências longas e de capturar relações mais complexas entre elementos em uma sequência, atingindo o que, segundo os autores Vaswani et al. (2017), seria "uma pequena fração dos custos de treinamento dos melhores modelos da literatura".

A transição da abordagem dos Transformers para o modelo BERT (Bidirectional Encoder Representations from Transformers) marcou um avanço significativo no processamento de linguagem natural (Devlin et al., 2019). Enquanto os Transformers introduziram a potência dos mecanismos de atenção para capturar relações de longo alcance em sequências, o BERT foi além ao aprimorar essa capacidade com o treinamento bidirecional de contextos, compreendendo o significado e o contexto de palavras em maior escala.

Devido à sua capacidade de capturar contextos amplos e complexos, Devlin et al. (2019) também demonstra em seu trabalho o poder do BERT ao ser aplicado com sucesso em diversas tarefas de NLP utilizando técnicas de *transfer-learning*. O diferencial estava em pré-treinar o modelo em grandes quantidades de texto não rotulado na internet, permitindo que o BERT captu-

rasse nuances linguísticas e semânticas, tornando-o uma base para "transferir seu conhecimento"(*fine-tuning*) para realização de tarefas de processamento de linguagem natural, como classificação de texto, perguntas e respostas, tradução e sumarização de textos.

Com essa revolução no PLN proporcionado por esses estudos, novas variações do BERT também surgiram em trabalhos com o intuinto de aprimorar ainda mais suas capacidades. (Liu et al., 2019), com RoBERTa, otimiza de forma robusta os hiper-parâmetros e tempo de treinamento do modelo. Enquanto DistilBERT (Sanh et al., 2019) e ALBERT (Lan et al., 2019) tentam criar versões mais compactas e ainda assim eficientes. Além de muitos outros modelos lançados cada qual com um propósito mais específico, porém ainda baseados no BERT (Sun et al., 2020; Yang et al., 2019; Clark et al., 2020; Conneau et al., 2019).

Cabe destacar aqui um modelo do BERT adaptado para o português brasileiro, chamado de BERTimbau (Souza et al., 2020). Esse LLM foi treinado com o propósito de realizar tarefas de PLN para um novo idioma em uma grande base da web brasileira como corpus (BrWaC - Brazilian Web as Corpus) (Wagner Filho et al., 2018). Além de demonstrar que o pré-treinamento em larga escala pode ser aplicado com sucesso a idiomas além do inglês, melhorando a compreensão de linguagem natural em português, ele provou ter melhor desempenho que o BERT multilíngue, confirmando sua efetividade.

3
Trabalhos Relacionados

Neste Capítulo, abordamos os trabalhos relacionados ao nosso estudo, que se concentra na investigação de técnicas associadas a sumarização de textos longos. Compreender como esses trabalhos evoluíram na criação de modelos, criação de corpus, assim como suas avaliações são importantes para gerarmos resumos de qualidade em português. Neste contexto, revisamos os estudos prévios que exploraram essas técnicas, analisando os principais avanços, desafios e lacunas na literatura científica existente.

3.1
Desenvolvimento dos modelos

As redes LSTM foram as primeiras redes neurais criadas para guardar informações de longo prazo em relações sequenciais como em textos, por exemplo. Yasunaga et al. (2019) faz uso dessa característica para criar resumos híbridos, que combinam abstração e extração de informações. Além de criar um extenso corpus anotado para sumarização de artigos científicos, os autores desenvolveram um modelo LSTM que utiliza um grafo de citações, grafo que conecta os artigos do corpus de acordo com citações e referências para dar mais contexto semântico ao gerar resumos que são uma mistura de abstração e extração, se aproximando mais dos textos escritos por humanos. Da mesma forma, An et al. (2021) propuseram uma sumarização aprimorada, que também envolve a criação de um grafo para utilizar referências prévias associadas a um modelo LSTM, gerando assim textos resumidos mais concisos. Por outro lado, Cohan et al. (2018) concentra-se especificamente na sumarização de artigos científicos sem o uso de referências, criando resumos unicamente a partir do texto.

Seguindo a tendência de mudança nas redes neurais, novos trabalhos de sumarização extrativa ou abstrativa de textos surgiram a fim de explorar as capacidades do *fine-tuning* do BERT. Liu (2019) investiga a sumarização extrativa, explorando o papel das camadas de Transformers na arquitetura do BERT, usando datasets de notícias (CNN/Dailymail e NY Times). Similarmente, (Srikanth et al., 2020) combina técnicas de extração e clusterização de video-aulas para gerar resumos extrativos sobre esses conteúdos para ajudar estudantes.

Por outro lado, Zolotareva et al. (2020), por exemplo, fez uso do BERT no contexto específico de notícias para criar resumos abstrativos curtos utilizando o dataset da BBC (Narayan et al., 2018). Seus resultados demonstram que o modelo ajustado superou os tradicionais de até então em tarefas de sumarização. Ma et al. (2021) propõe o uso do BERT considerando o tema associado a partir de uma base externa a fim de criar resumos abstrativos com mais representação contextual.

No mesmo sentido, vale destacar o trabalho realizado por Liu e Lapata (2019) também propõe um modelo que realiza o *fine-tuning* do BERT como parte de um *encoder* para alcançar um bom desempenho de sumarização. Os autores propõem um modelo mais simples e que possui menos parâmetros, o que significa precisar de menos recursos de hardware em comparação com outros trabalhos de ponta na literatura. Os resultados mostram que, mesmo com menor tamanho, esse modelo obteve um bom desempenho ao realizar tarefas de sumarização extrativa e abstrativa para criar resumos de artigos de notícias em diferentes conjuntos de dados.

No que se refere a trabalhos com aplicação focada na língua portuguesa, Paiola et al. (2022) traz o estudo mais recente para a sumarização de textos curtos de notícias utilizando técnica de *fine-tuning* do modelo pre-treinado PTT5 (Carmo et al., 2020), um LLM para português similar ao BERTimbau, traduzindo seus textos de português para inglês e, em seguida, sumarizando com o objetivo de comparar seus resultados ao desempenho de outros trabalhos em inglês. A ideia é que, ao traduzir os textos para o inglês e, em seguida, sumarizá-los, eles possam obter sumarizações comparáveis e avaliar a qualidade das sumarizações geradas pelo modelo T5, na falta de outros *baselines* em português.

Paiola et al. (2022) cita outros estudos que tratam de sumarização abstrativa automática usando outras abordagens como: compressão de sentenças (Martins e Smith, 2009; Nóbrega e Pardo, 2016), que envolve a redução do texto original, eliminando palavras, frases ou partes de sentenças que são consideradas menos relevantes para a essência do conteúdo; ou usando AMR (Abstract Meaning Representation), realizada considerando a estrutura semântica subjacente do texto. O AMR é um formato que representa o significado de frases e sentenças em uma forma mais lógica e independente do idioma. (Inácio, 2021).

3.2 Domínio e corpus

Nós buscamos datasets em português que suprissem nossa necessidade: um conjunto de textos longos anotados com resumos sobre a área da saúde. No

entanto, esses dados específicos são escassos e tivemos que construir um. Sendo assim, como construir um corpus relevante? Nós, então, buscamos na literatura em inglês como são os corpora para sumarização de texto que servissem de referência.

Existem diversos domínios de pesquisa na língua inglesa, cada um com suas características únicas. Muitos estudos têm se concentrado na sumarização de artigos de notícias, beneficiando-se de conjuntos de dados fornecidos por grandes agências de imprensa e anotados por pesquisadores. Exemplos desses conjuntos incluem o **New York Times** (cerca de 100 mil artigos e seus resumos)(Durrett et al., 2016), **CNN/DailyMail** (mais de 300 mil exemplos) (Hermann et al., 2015) e **BBC** (cerca de 220 mil) (Narayan et al., 2018). Além disso, conjuntos de dados como **NEWSROOM** (1,3 milhão de textos) (Grusky et al., 2018), **Multi-News** (mais de 250 mil) (Fabbri et al., 2019) e **Gigaword** (por volta de 4 milhões de artigos) (Graff et al., 2003) têm sido utilizados para avaliação de soluções de PLN.

Diferentes domínios também têm sido explorados para a sumarização, com conjuntos de dados abrangendo tópicos diversos. Por exemplo, o **BIGPATENT** foi criado a partir de patentes dos EUA com 1,3 milhão de resumos abstrativos escritos por humanos (Sharma et al., 2019), enquanto o **BillSum** contém pouco mais de 22 mil projetos de lei do congresso dos EUA com resumos de sessões (Kornilova e Eidelman, 2019). Além disso, comunidades online como o Reddit fornecem conjuntos de dados, como o **Reddit TIFU**, com mais de 120 mil postagens e seus resumos TLDR (Kim et al., 2018), apesar da possibilidade de conter linguagem informal e vocabulário específico sobre algum assunto.

A sumarização de artigos científicos também é abordada com a criação de conjuntos de dados que contêm citações, facilitando o acesso ao conteúdo das referências. Exemplos incluem o **Semantic Scholar Network (SSN)** contendo 141 mil artigos, com mais de 660 mil relacionamentos entre eles (An et al., 2021) e o **ScisummNet**, que contém aproximadamente mil resumos anotados manualmente, sendo que em alguns casos, certos textos possuem mais de um resumo associado (Yasunaga et al., 2019). Outros conjuntos de dados, como o **CL-SciSumm**, **ACL Anthology Network (AAN)**, **Microsoft dataset**, **cmp-lg corpus** e o **PLOS8 Medicine corpus**, também contribuem com artigos científicos em diferentes tópicos e estruturas para a conexão entre as referências.

Na língua portuguesa, destacamos alguns datasets em domínio diversos, visando resolver o desafio da falta de dados em Português. Para abordar essa questão mencionamos os seguintes conjuntos:

- o **RulingBR** contém cerca de 10 mil textos completo e seus resumos jurídicos com decisões do Supremo Tribunal Federal (STF) brasileiro (de Vargas Feijó e Moreira, 2018);
- **CSTNews** contém 195 artigos de notícias a partir de Folha de São Paulo, Estadão, O Globo, Jornal do Brasil e Gazeta do Povo, manualmente anotados (Aleixo e Pardo, 2008)
- **Temário** contém cerca de 100 artigos da Folha de São Paulo e Jornal Brasil manualmente anotados
- **XL-Sum** é um dataset multilíngue com mais de 1 milhão de textos de notícias da BBC em 44 idiomas (Hasan et al., 2021), incluindo mais de 23 mil em português
- **Wikilingua** que contém cerca de 770 mil pares de texto e resumo, sendo mais de 81 mil em português (Ladhak et al., 2020)

Estes conjuntos têm como propósito fornecer dados anotados adequados para uma variedade de tipos de documentos e para a realização de resumos automáticos de texto, que consistem em gerar resumos breves a partir de textos extensos. Apesar de oferecerem essa riqueza de dados, o fato de estarem direcionados para textos mais curtos não se aplica ao nosso contexto, uma vez que nosso objetivo é resumir textos longos. Isso ressalta o papel da nossa coleção de documentos como uma contribuição para a área tanto a nível de quantidade de documentos quanto por se tratar de documentos longos.

3.3 Métodos de avaliação

Para efetuar a avaliação do desempenho de um modelo voltado à geração textual, é imperativo que um modelo avaliativo considere não apenas quão bem escrito, mas também a coesão e precisão com relação ao conteúdo do resumo gerado. Esse processo é complexo, particularmente no contexto da sumarização abstrativa, a qual pode apresentar discrepâncias em relação ao *target* predefinido durante a fase de treinamento. Assim sendo, a avaliação requer uma abordagem criteriosa para abarcar essas nuances.

Ibrahim Altmami e El Bachir Menai (2022) postula que as técnicas de avaliação podem ser categorizadas em intrínsecas e extrínsecas. A abordagem intrínseca mensura a qualidade de um sumário, considerando fatores como a integridade das sentenças, legibilidade, relevância, abrangência e exatidão, quando comparados a um padrão-ouro (um alvo ótimo, por assim dizer). Nesse cenário, sumários abstrativos podem enfrentar avaliações desfavoráveis, dado que possuem a capacidade de reformular frases e estruturas considerando ainda

a potencial geração de diversos sumários empregando vocabulários distintos. A avaliação extrínseca, por sua vez, contempla critérios diversos, como tempo de conclusão, taxa de sucesso e acurácia na tomada de decisões.

Em um dos primeiros trabalhos a tratar da avaliação automática de textos, Papineni et al. (2002) propuseram o BLEU (*Bilingual Evaluation Understudy*) como um método automático de avaliação para a tradução automática, a qual se insere no espectro das técnicas generativas. O BLEU foi concebido com o propósito de capturar as nuances nas escolhas efetuadas pelo modelo durante o processo de geração textual, empregando uma métrica de precisão de n-grama modificada. Entretanto, estudos subsequentes evidenciaram que o BLEU não guarda correlação substancial com a avaliação humana, sendo assim considerado como inadequado para a comparação entre sistemas dessa natureza (Callison-Burch et al., 2006). Dentre os estudos científicos examinados neste contexto, destaca-se a utilização do BLEU para avaliar pesquisas em Raffel et al. (2020); An et al. (2021).

De forma similar, surgiu o ROUGE (Recall-Oriented Understudy for Gisting Evaluation), desenvolvido por Lin (2004), que consiste em um método automático para avaliar a similaridade entre sumários. Embora tenha sido proposto há bastante tempo, em comparação com os padrões de tecnologia mais recentes, o ROUGE permanece como a métrica de avaliação automática padrão na maioria das publicações recentes (Fabbri et al., 2021).

Dentre as quatro distintas medidas advindas do ROUGE - ROUGE-N, ROUGE-L, ROUGE-W e ROUGE-S - destaca-se a ROUGE-N, tal como delineada nos estudos de Lin (2004), a qual representa uma "recuperação de n-grama entre um sumário candidato e um conjunto de sumários de referência". Seu objetivo consiste em mensurar a proximidade entre um sumário candidato e o consenso presente nas múltiplas referências (*target*). Em paralelo, a métrica ROUGE-L quantifica as subsequências comuns mais longas. Por sua vez, a métrica ROUGE-W aborda a subsequência comum mais longa ponderada. A métrica ROUGE-S, a última dentre as quatro, diz respeito às Estatísticas de Co-Ocorrência de Bigramas Ignorados, permitindo a contabilização da ordem arbitrária das sentenças. A métrica ROUGE revela-se presente nos estudos de Raffel et al. (2020); An et al. (2021); Zolotareva et al. (2020); Liu e Lapata (2019); Yasunaga et al. (2019); Xiao e Carenini (2019); Cohan et al. (2018).

De modo abrangente, os resultados obtidos requerem a adoção de mais de um método avaliativo, conduzindo pesquisadores a empregar ocasionalmente mais de uma única métrica, com o intuito de contemplar outros pontos de vista para a análise. A busca por uma análise abrangente é corroborada por métricas como acurácia, *F1-score*, *recall*, precisão e outros métodos. Ibrahim Altmami e

Referência	**Método**	**Corpus**	**Avaliação automática**
Saggion e Lapalme (2000)	Extrativo	Autoria própria	Recall, Precision e F-Score
Lloret et al. (2013)	Extrativo e Abstrativo	Autoria própria	ROUGE-1
Yang et al. (2016)	Extrativo	AAN, Microsoft dataset	ROUGE-1, ROUGE-2
Slamet et al. (2018)	Extrativo	Autoria própria	Manual

Tabela 3.1: Ranking dos 10 unigramas e bigramas mais frequentes nos textos principais do corpus.

El Bachir Menai (2022) proporciona uma visão das abordagens utilizadas por outros trabalhos para a avaliação automática de suas experimentações para a geração automática de resumos, conforme retratado na Tabela 3.1.

Além disso, alguns estudos têm se dedicado à condução de avaliações qualitativas sobre os resumos abstrativos gerados pelos modelos propostos. Por exemplo, Saggion e Lapalme (2000) conduziu um experimento em que avaliadores individuais foram convidados a ler sentenças geradas pelos modelos e classificá-las como aceitáveis ou não. Enquanto isso, Wang et al. (2017) adotaram uma abordagem diferente, recorrendo a dois avaliadores humanos para atribuir notas em uma escala de 5 pontos aos resumos produzidos. Essa abordagem visava avaliar a qualidade dos resumos sem revelar as referências originais aos avaliadores. Por fim, Yasunaga et al. (2019) também empregaram uma escala de 5 pontos para que os participantes avaliassem especificamente a coerência e a cobertura dos resumos gerados em relação aos resumos de referência.

4
Construção do Corpus

Neste Capítulo, apresentaremos a metodologia adotada para solucionar o problema da sumarização de textos. Como apresentado no Capítulo anterior (Cap. 2), diversos estudos têm se aprofundado na busca por soluções eficazes, dedicando-se à construção de experimentos sólidos e bem fundamentados, resultando em um conjunto abrangente de pesquisas presentes na literatura. Sendo assim, o enfoque de nossa pesquisa se desdobra em três etapas fundamentais: (a) a construção de um corpus apropriado, (b) o desenvolvimento e treinamento de modelos de aprendizado de máquina voltados à sumarização e (c) a avaliação dos resultados obtidos. Cada uma dessas etapas representa um pilar na estruturação do nosso estudo, contribuindo para uma abordagem abrangente na análise posterior dos resumos obtidos, também representada na Figura 4.1.

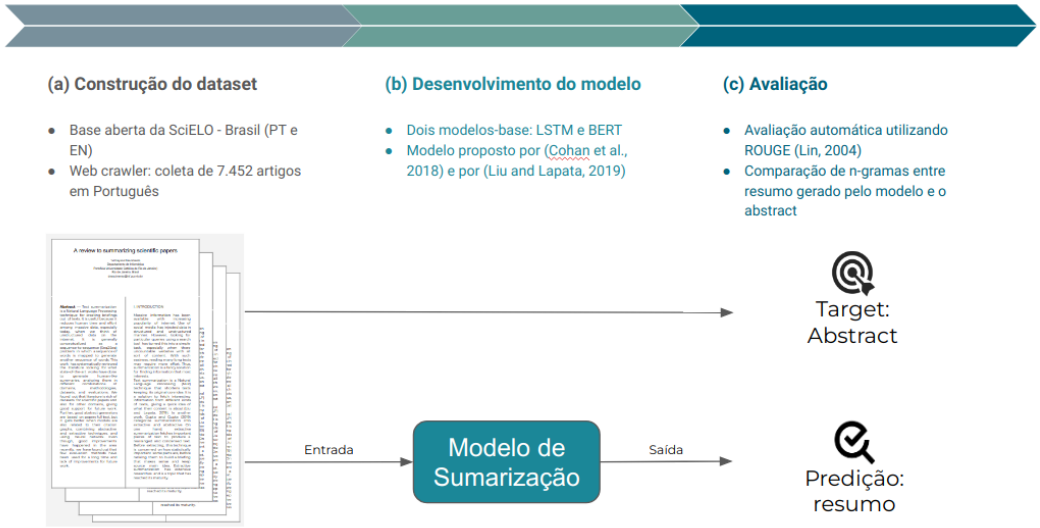


Figura 4.1: Metodologia utilizada nesse trabalho.

Cada uma dessas etapas serão detalhadas nas seções seguintes.

4.1
Coleta de dados

A escassez de conjuntos de dados na língua portuguesa, especialmente aqueles que abrangem textos mais extensos, tem sido uma limitação enfrentada por

outros estudos (Nóbrega e Pardo, 2016; Paiola et al., 2022). Para viabilizar o nosso trabalho, buscamos enfrentar essa lacuna criando um corpus composto exclusivamente por artigos científicos no domínio das Ciências da Saúde escritos em português. Vale ressaltar que esse documentos são públicos e acessíveis pelo portal do SciELO-BR[1]

Além de utilizar esse corpus para treinar os algoritmos empregados neste estudo, também esperamos que este também sirva como material de suporte para outros futuros trabalhos na área. Para garantir esse suporte, o corpus está disponível em repositório público[2]. A Seção seguinte apresenta todo o processo de aquisição desses documentos.

4.1.1 Web Crawler

Para a construção desse corpus, utilizamos a plataforma SciELO-BR[3]. A SciELO, ou Biblioteca Eletrônica Científica Online, é uma biblioteca digital de livre acesso e um projeto cooperativo de publicação digital de periódicos científicos. Foi criada em 1997 em São Paulo, Brasil, em colaboração com a FAPESP e a BIREME, recebe apoio do CNPq desde 2002, com o objetivo de ampliar o acesso ao conhecimento.

O foco do corpus recaiu sobre o campo das Ciências da Saúde, sendo estabelecido que dos documentos filtrados seguiriam os seguintes critérios:

i documento deveria conter mais de 500 palavras, o que seria equivale a limitar à artigos com mais de uma página de conteúdo textual;

ii resumo deveria conter de 180 a 300 palavras;

iii versão em HTML disponível, descartando os artigos em PDF ou Latex; e

iv textos deveriam estar escritos em português.

A coleta dos dados foi realizada por meio do desenvolvimento de um *web crawler*, um pequeno software que automatizou a coleta de documentos[4]. Um software versátil que pode ser configurado para buscar em diferentes domínios da base de dados SciELO-BR com pequenas modificações, o que oferece potencial para expandir ainda mais o corpus.

[1] Acessível em `https://www.scielo.br/`. Último acesso em 11/10/2022.

[2] `https://drive.google.com/drive/folders/1165NTu8pbern-xPNv-vFV9_m0ZVe1XzD?usp=sharing`

[3] SciElo-BR: `https://www.scielo.br/`

[4] O código-fonte desse *web crawler* foi tornado acessível publicamente por meio de um repositório no GitHub (`https://github.com/daysonn/papers_crawler`) permitindo que outros possam utilizá-lo e adaptá-lo facilmente.

O critério de filtragem de artigos em HTML possibilitou uma segmentação mais adequada das seções dos artigos para ser usado posteriormente. O formato HTML também permite uma extração precisa dos dados, uma vez que elementos comuns em artigos científicos, como tabelas, figuras e equações, possuem uma estrutura bem definida e separada no HTML em si. Essas escolhas nos permitem ter dados confiáveis e de boa qualidade para nossa pesquisa, garantindo que estamos analisando os artigos de forma correta.

Como resultado dessa coleta utilizando o *web crawler*, coletamos um conjunto de 11.416 documentos em HTML. Após isso, seguimos uma etapa de refinamento e exclusão, no qual esse conjunto foi reduzido para 7.452 artigos científicos. Para esse processo usamos os seguintes critérios de exclusão:

i Textos que contêm apenas links para PDFs

ii Erratas, análises ou outros textos sem resumos

iii Textos duplicados

4.1.2 Tratamento de dados

Na etapa do pré-processamento dos dados, realizamos um tratamento com base na análise das palavras e seções. Essa etapa combinou abordagens da literatura que guiam nosso trabalho (Cohan et al., 2018; Liu e Lapata, 2019). Nessa etapa, os seguintes critérios foram utilizados:

- Remoção de figuras, tabelas, equações e outros elementos não textuais
- Remoção de *key words*, agradecimentos, bibliografia e colaboradores
- Transformação do texto para *lower case*
- Padronização de números inteiros e decimais para formato brasileiro
- Espaçamento da pontuação para facilitar tokenização
- Separação entre resumo e o restante do documento

Essa fase de tratamento dos dados resultou em uma padronização mais consistente do conteúdo de cada artigo. Como resultado, os artigos processados passaram a ter em média 3.951 palavras (aumento de 4,5% em virtude da exclusão de 34,7% dos artigos inicialmente coletados) em sua forma completa, enquanto os resumos correspondentes têm, em média, 180 palavras. Essas etapas nos proporcionam uma base de dados mais ajustada para sequência em nossa pesquisa.

4.2 Análise do Corpus

Com base nesse contexto, realizamos uma análise exploratória dos dados contidos em nosso corpus, composto por uma coleção de 7.450 artigos científicos redigidos em língua portuguesa e um vocabulário de pouco mais de 240 mil palavras. Nessa investigação, nosso propósito consistiu em compreender bem os nossos dados e assim validar as informações e soluções do nosso trabalho, assim como sugerido por Kalinowski et al. (2023), ao abordar a análise exploratória de dados.

Classe de palavras	**% de tipos**
Substantivo	27,79
Preposição	18,80
Pontuação	13,02
Verbo	10,07
Adjetivo	8,70
Artigos	7,08
Conjunção	4,53
Número	4,00
Nome próprio	2,02
Advérbio	1,85
Pronome	1,04
Interjeição	0,01

Tabela 4.1: Análise morfológica das palavras contidas no dataset.

Uma demonstração da análise morfológica dos elementos contidos nesse corpus nos mostra a distribuição geral das classes de palavras ao longo dos artigos científicos, detalhada na Tabela 4.1. Algumas classes não foram incluídas por representarem uma parte muito pequena do percentual total.

Além disso, os artigos coletados foram publicados entre os anos de 2000 a 2022. A Tabela 4.2, apresenta um descrição detalhada do corpus por ano de publicação:

Nas seções a seguir, buscamos compreender melhor tanto resumos quanto textos principais dos artigos coletados. Essa abordagem busca fornecer informações mais sólidas sobre os dados presentes em nosso corpus como: palavras e bigramas mais frequentes, além da distribuição da quantidade de palavras em um histograma.

4.2.1 Descrição dos Resumos

A etapa inicial de nossa análise concentrou-se na identificação das palavras mais frequentemente utilizadas em nossa coleção, a qual se concentra no âmbito das Ciências da Saúde. Especificamente, direcionamos nossa atenção

Ano	Total de Artigos	Tokens		
		Total	Média	Std.
2000	2	3.251	1.625,50	375,47
2001	284	924.668	3.255,89	1.394,42
2002	183	640.688	3.501,03	1.428,70
2003	228	736.364	3.229,67	1.542,79
2004	307	1.057.083	3.443,27	1.455,72
2005	270	860.114	3.185,61	1.457,10
2006	558	1.953.060	3.500,10	1.455,21
2007	461	1.594.574	3.458,95	1.299,31
2008	395	1.301.209	3.294,20	1.386,85
2009	372	1.296.971	3.486,48	1.323,90
2010	472	1.481.416	3.138,59	1.332,67
2011	402	1.289.430	3.207,54	1.295,29
2012	453	1.415.576	3.124,89	1.292,01
2013	330	961.482	2.913,58	1.241,04
2014	315	1.346.038	4.273,14	1.442,81
2015	277	1.322.325	4.773,74	1.315,31
2016	262	1.354.270	5.168,97	1.243,19
2017	296	1.563.283	5.281,36	1.172,79
2018	274	1.460.756	5.331,23	1.186,89
2019	303	1.594.122	5.261,13	1.123,77
2020	344	1.776.137	5.163,19	1.181,59
2021	353	1.887.610	5.347,34	1.075,53
2022	302	1.620.559	5.366,09	1.106,21
Total	7.450	29.440.165	3951,70	1266,50

Tabela 4.2: Mais detalhes sobre o corpus

aos resumos, que desempenharão o papel de alvos (*targets*) para os modelos submetidos a treinamento. Tais resumos são representados em uma nuvem de palavras, como ilustrado na Figura 4.2.

Além disso, apresentamos uma visão complementar dessa análise na Tabela 4.3, onde destacamos os dez unigramas que mais aparecem em nosso corpus. A palavra "saúde"é a mais recorrente no corpus, o que está coerente com área de domínio dos artigos. Vale destacar a presença de palavras como "estudo", "resultado"ou "resultados", "dado"ou "dados"e "análise", que apontam para uma ênfase em atividades de análise de dados nesses trabalhos. Essa descoberta está alinhada com a natureza intrínseca do método científico, uma vez que essas expressões são consistentes com a abordagem investigativa típica do campo.

A análise dos bigramas mais frequentes (também incluídos na Tabela 4.3) nos ajuda a compreender como as palavras são usadas em pares e a identificar padrões recorrentes de expressões. Nessa análise, excluímos *stopwords*, para nos concentrarmos nas palavras mais relevante. Nessa Tabela, cabe destacar algumas tendências no estilo de escrita. Por exemplo, é possível verificar que os autores resumiram seus estudos com uma estrutura lógica que pode incluir objetivos, metodologia e conclusão.

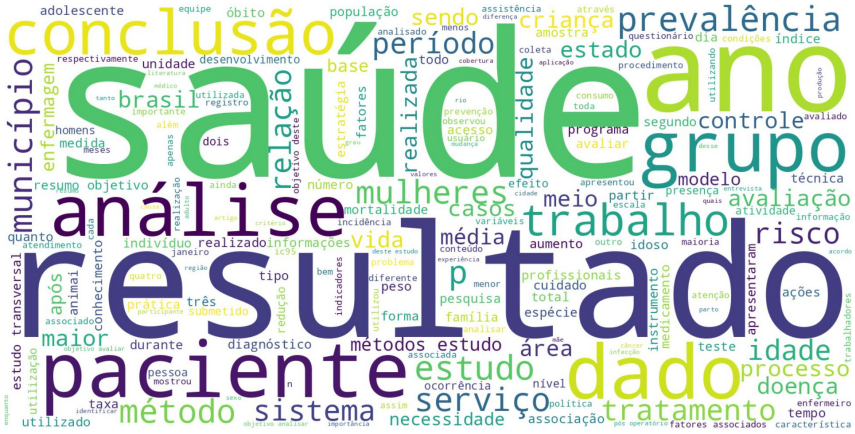


Figura 4.2: Nuvem de palavras com as palavras mais frequentes dos resumos contidos no corpus.

Ranking	Unigrama		Bigrama	
1	Saúde	8.060	Métodos estudo	1.299
2	Resultado	4.707	Resumo objetivo	1.259
3	Ano	4.110	Estudo transversal	860
4	Paciente	3.522	Fatores associados	632
5	Dado	3.345	Objetivo deste	614
6	Análise	3.022	Objetivo avaliar	601
7	Grupo	2.934	Objetivo analisar	576
8	Conclusão	2.627	Qualidade vida	548
9	Trabalho	2.579	Rio janeiro	541
1110	Prevalência	2.418	Serviços saúde	521

Tabela 4.3: Ranking dos 10 unigramas e bigramas mais frequentes nos resumos do corpus.

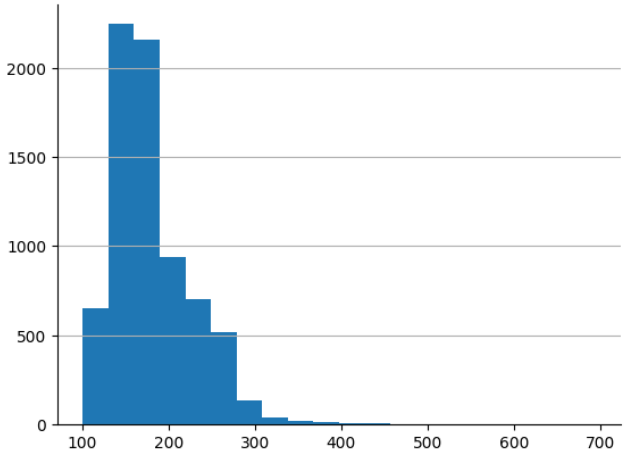


Figura 4.3: Histograma do tamanho dos resumos.

Como mencionado anteriormente, os resumos dos artigos possuem uma média de 181 palavras. A Figura 4.3 ilustra a distribuição de tamanho dos resumos. Vale observar que a extensão dos resumos varia dentro de uma faixa que vai de 100 a 694 palavras, refletindo uma alta variação nas quantidades de

Ranking	Unigrama		Bigrama	
1	Saúde	162.432	Ministério saúde	17.782
2	Ano	78.873	p< 0,05	13.195
3	Estudo	65.839	Saúde pública	11.907
4	Paciente	63.656	Pode ser	10.240
5	Dado	59.752	Rio janeiro	9.996
6	Trabalho	59.507	Serviços saúde	9.338
7	Relação	57.928	Presente estudo	7.153
8	Serviço	56.634	Faixa etária	6.287
9	Análise	49.415	Além disso	6.081
10	Município	42.409	Pré natal	5.968

Tabela 4.4: Ranking dos 10 unigramas e bigramas mais frequentes nos textos principais do corpus.

palavras, tornando a tarefa de sumarização ainda mais desafiadora, por não haver um padrão claro entre os autores (qualquer tamanho é válido desde que seja menor que 700 palavras).

4.2.2 Descrição dos Textos Principais

De maneira semelhante, realizamos uma análise focalizada no corpo principal dos artigos científicos, que abrange toda a apresentação da pesquisa e exclui o resumo. Da mesma forma que na Subseção 4.2.1, apresentamos uma representação visual dos termos mais frequentes ao longo desse conteúdo, na Figura 4.4, além de uma tabela detalhada das dez palavras mais utilizadas na Tabela 4.4.

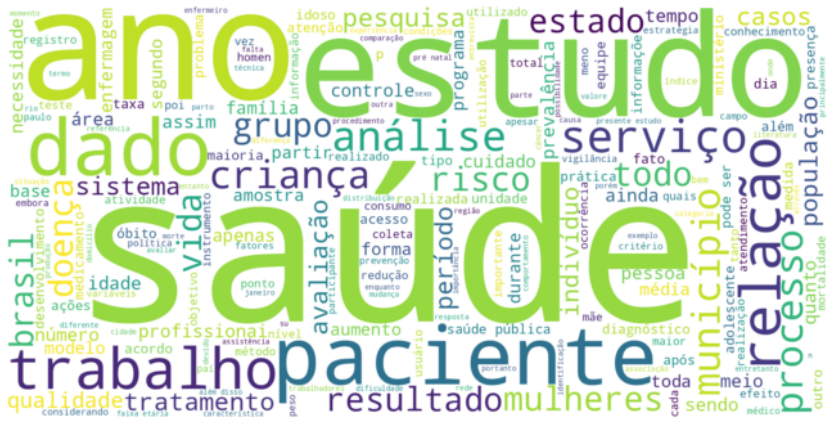


Figura 4.4: Nuvem de palavras com as palavras mais frequentes dos textos principais dos artigos.

Vale ressaltar que a Tabela 4.4 proporciona uma visão alternativa da estrutura dos textos principais: a presença da indicação "p<0,05" sugere a existência de um viés analítico nos textos; outros bigramas, como "ministério saúde" e "saúde pública", apontam para informações de acesso público, pro-

vavelmente provenientes de fontes governamentais. Isso também sugere que, durante o treinamento dos modelos, os números presentes nas pesquisas não podem ser omitidos ou mascarados, e a precisão dos modelos depende da seleção adequada desses números nos textos que irão compor os resumos gerados.

No que diz respeito ao comprimento dos textos, o corpo principal tem uma média de 3.210 palavras (algo em torno de 8 páginas), o que destaca sua extensão considerável como entrada para nossos modelos. Isso evidencia que essas seções são substanciais em termos de conteúdo, o que é relevante, pois elas são usadas como entrada nos nossos modelos. Além disso, essa extensão varia consideravelmente, com um mínimo de 551 palavras e um máximo de 5.581 palavras (de 2 a 15 páginas). Isso sugere que os textos abrangem um amplo espectro de comprimentos, o que pode refletir a diversidade e complexidade das pesquisas abordadas, mantendo-se em consonância com o objetivo do nosso trabalho de sumarizar textos longos.

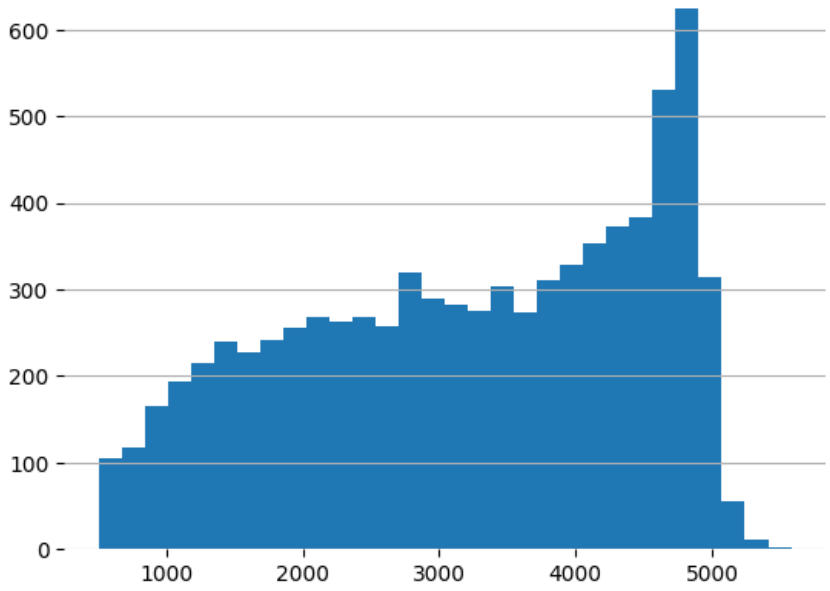


Figura 4.5: Histograma tamanho do corpo principal dos artigos.

O histograma representado na Figura 4.5 oferece uma visualização da distribuição do tamanho dos artigos científicos. Esse gráfico nos permite entender como os diferentes comprimentos dos textos estão distribuídos dentro da nossa coleção. Quanto à estrutura dos artigos, eles, em média, são compostos por cerca de 3,8 seções. Essa medida ressalta a organização e a divisão temática que são características típicas dos artigos científicos.

5
Geração dos resumos

Como apresentado e discutido nos Capítulos anteriores, nosso estudo tem o foco em dois modelos para geração de resumos: o primeiro, proposto por Cohan et al. (2018), e o segundo proposto por Liu e Lapata (2019). O primeiro, está em consonância com nosso objetivo de resumir textos longos sem depender de grafos de referência. Faremos uma comparação entre este modelo e aquele proposto por Liu e Lapata (2019), mais recente, que resume textos a partir do *fine-tuning* do BERT (Devlin et al., 2019). Tanto um quanto o outro modelo representam abordagens de ponta, oferecem repositórios publicamente disponíveis para garantir a replicabilidade e utilizam arquiteturas distintas de redes neurais. Isso nos proporciona uma oportunidade para avaliar seu desempenho quando aplicados ao contexto da língua portuguesa.

Nas próximas seções iremos descrever a metodologia aplicada na nossa proposta de processo de treino para cada um dos modelos. Vale ressaltar que ambos utilizaram para a etapa de treino 5960 (80%) dos artigos presentes no copus, 750 (10%) para teste e 750 (10%) para validação.

5.1
LSTM *from the scratch*

Com o objetivo de servir como *baseline* de comparação, uma referência de resultado de qualidade mínima. O primeiro modelo foi desenvolvido baseado em LSTMs e o seu treinamento foi feito a partir do zero. Esta proposta de modelo foi a mesma apresentada por Cohan et al. (2018), que a utilizou para o objetivo especifico de resumir artigos científicos escritos em inglês.

Esse modelo faz uso do *framework* sequência para sequência (*Sequence-to-Sequence* ou Seq2Seq) em que um *encoder* recebe todo o documento como entrada e transforma o texto em uma representação vetorial para que, em seguida, um *decoder* seja utilizado para decodificar aquela representação em um resumo. No entanto, Cohan et al. (2018) propõe que *encoder* e *decoder* mantenham a semântica das seções distribuídas ao longo de um documento, preservando, assim, o conteúdo em textos mais extensos.

O modelo inicia o processo com o *encoder*, que é responsável por analisar e codificar cada seção do documento de forma separada, utilizando uma LSTM

para realizar essa tarefa. Para cada seção do documento, o *encoder* gera uma matriz de pesos. Essa matriz de pesos passa, então, a refletir a estrutura e importância dessa seção específica no contexto do documento.

A partir daí, as matrizes de pesos geradas para todas as seções são passadas para outra camada de LSTM. Essa segunda camada de LSTM tem a função de compilar todas as entrada das diferentes seções e criar uma representação global e coesa do texto completo. Ela integra as informações relevantes de todas as seções de forma a capturar a essência do conteúdo.

Portanto, o *encoder* opera em duas etapas: primeiro, codifica cada seção individual do documento usando uma rede LSTM para gerar matrizes de pesos que refletem a estrutura de cada seção. Em seguida, essas matrizes de pesos são processadas por outra camada LSTM para criar uma representação textual completa e enriquecida que reflete as nuances e relações entre as diferentes partes do documento.

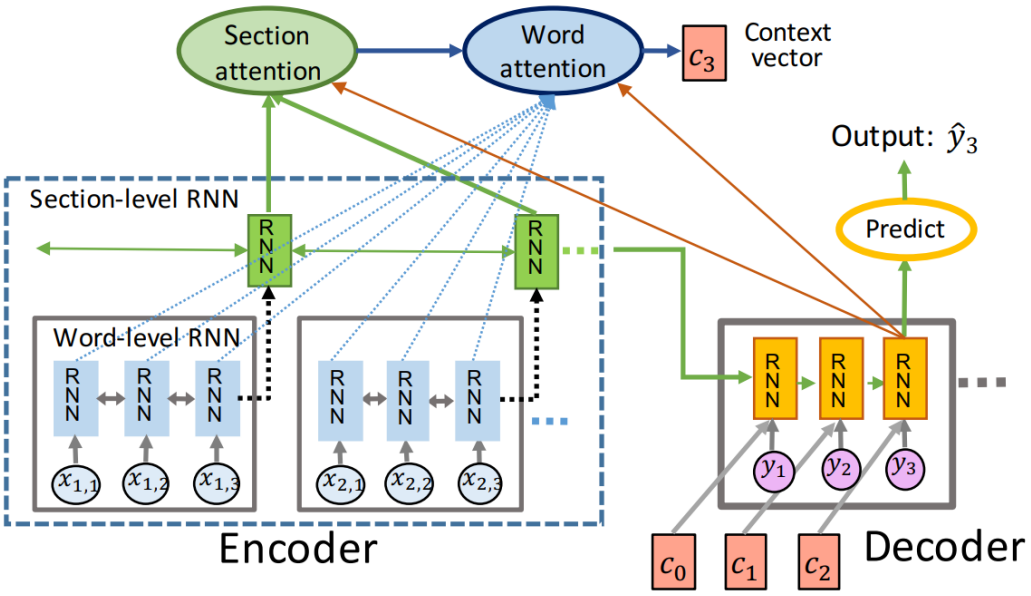


Figura 5.1: Visão geral do modelo de Cohan et al. (2018). Representação do *step* 3 de decodificação. A menção a RNN sempre se refere às LSTMs Bidirecionais.

Na outra ponta, temos o *decoder* deste modelo, que também é construído com redes LSTM. Ele recebe a saída produzida pelo *encoder* e realiza predições, que consistem em palavras para formar o resumo. Cada *step* passa por essa rede LSTM e gera uma nova palavra, que por sua vez contribui para compor o resumo final. Isso ocorre até alcançar um limite predefinido de palavras no resumo ou até que a rede gere um sinal (em geral um *token*) indicando o final de uma frase.

O que torna o trabalho de Cohan et al. (2018) especial é a adição de duas camadas de "atenção" para compor o discurso final. A primeira camada de

atenção se concentra em calcular pesos para as diferentes seções do documento. Isso significa que, em vez de tratar todas as partes do texto igualmente, a rede pode dar mais importância a certas seções que são consideradas mais relevantes para a geração do resumo. A segunda camada de atenção opera em um nível mais granular: no nível das palavras individuais. Ela calcula pesos para as palavras dentro de cada seção. Isso permite que o modelo selecione palavras-chave e informações cruciais em cada parte do texto, para que esses elementos sejam mais bem representados no resumo. Essas duas camadas de atenção são combinadas, resultando em um vetor de contexto. Esse vetor de contexto é então utilizado para orientar cada passo de predição realizado pelo *decoder*.

Dessa forma, o *decoder* não apenas gera palavras do resumo, mas faz isso de maneira aprimorada, graças à influência das camadas de atenção. Elas permitem que o modelo focalize nas seções e palavras mais relevantes, criando um resumo mais preciso e informativo do texto original.

A Figura 5.1 nos dá uma ideia geral de como o modelo proposto por Cohan et al. (2018) foi construído. Importante notar que, na imagem, o termo "RNN" é usado para se referir às LSTMs, que são uma versão aprimorada das primeiras redes neurais recorrentes. Essa figura resume a abordagem desenvolvida e mostra como o modelo funciona de maneira geral.

Os dados gerados a partir do nosso corpus estão divididos pelas seções de cada artigo e dentro de cada seção há uma divisão por sentenças para servirem como entrada para o modelo. A tokenização por sentença foi feita utilizando a biblioteca NLTK [1], uma ferramenta gratuita, *open source* e que fornece esse tipo de tarefa pronta para o português.

Com isso, nosso modelo foi treinado durante 7 dias para que pudesse resumir os longos artigos científicos. Da mesma forma que Cohan et al. (2018), utilizamos mini-batches de tamanho quatro e impusemos restrições ao comprimento do documento (3.000 tokens), ao comprimento da seção (600 tokens) e ao número de seções (5). Utilizamos o otimizador Adagrad com uma taxa de aprendizado de 0,15 e um valor inicial de acumulador de 0,1. O tamanho máximo do decodificador foi definido como 400 tokens, alinhando-se com o comprimento médio dos resumos do nosso corpus.

Vale ressaltar que nesse tempo, nosso experimento atingiu 20 mil interações enquanto Cohan et al. (2018) treinou seu modelo durante 250 mil interações. As diferenças entre os dois trabalhos estão no tempo de treino, no idioma e no tamanho dos corpus empregados. Portanto, não seria uma comparação justa entre os dois. Sendo assim, esse processo de treinamento torna-se mais comparável como *baseline* em relação ao modelo proposto que será apresentado

[1] NLTK: `https://www.nltk.org/`

na Seção 5.2.

5.2 *Fine-tuning* uma versão do *BERTimbau*

No segundo modelo utilizado nesse trabalho, exploramos o *fine-tuning* aplicado ao modelo BERT (Devlin et al., 2019), seguindo a metodologia proposta por (Liu e Lapata, 2019), que a utilizou para sumarizar textos em inglês. Para o nosso contexto, o modelo com o BERT foi adaptado para lidar com textos em Português.

Para isso, essa adaptação envolveu uma etapa diferente do trabalho original: substituir o BERT pelo modelo BERT pré-treinado para o português brasileiro, o BERTimbau (Souza et al., 2020). Seguindo o framework *encoder-decoder*, essa alteração implicou em utilizar o BERTimbau como *encoder* do texto completo assim como o proposto por (Liu e Lapata, 2019), antes de envolver um re-treinamento do modelo com nosso corpus em Português para a sumarização de textos longos.

Da mesma forma que o *decoder* mencionado na Subseção 5.1, o *decoder* deste modelo também opera gerando predições até atingir um *token* final ou atingir um limite máximo de palavras. No entanto, sua estrutura é composta por 6 camadas de *Transformers*, que são redes neurais desenvolvidas para compreender o contexto e os significados em relações sequenciais (Vaswani et al., 2017).

Vale ressaltar que, assim como mencionado por (Liu e Lapata, 2019), o BERTimbau está sendo utilizado apenas como parte do *encoder* e o *decoder* não é pré-treinado e, portanto, precisa de uma etapa de treinamento. Por conta disso, é possível que haja uma diferença entre o *encoder* e o *decoder*, já que o primeiro é pré-treinado e o segundo precisa ser treinado do zero. Isso pode tornar o *fine-tuning* instável, com um deles se ajustando demais e o outro de menos.

Para evitar isso, da mesma forma que (Liu e Lapata, 2019), utilizamos dois otimizadores Adagrad com duas taxas de aprendizado: uma mais baixa de $2e^{-3}$ para o *encoder* e outra mais alta de 0,1 para o decoder, assim, evitando que haja *overfitting* da parte pré-treinada ou *underfit* do *decoder*.

Após um período de treinamento de 7 dias, tanto o *encoder* (responsável por condensar as informações) quanto o *decoder* (responsável por gerar o resumo) dos modelos convergem para produzir resumos a partir dos artigos científicos completos. Esse processo de convergência ocorre ao longo de um total de 180 mil interações, que envolvem ajustes e refinamentos nos parâmetros do modelo.

É importante notar que o estudo conduzido por Liu e Lapata (2019) atingiu um total de 200 mil interações, fazendo uso de 4 unidades de processamento gráfico (GPUs). No entanto, é essencial destacar que as condições e os objetivos desse estudo podem diferir substancialmente dos nossos, tornando uma comparação direta entre os dois trabalhos difícil e potencialmente inválida.

Portanto, ao analisar e discutir os resultados deste estudo, devemos considerar as particularidades dos métodos empregados (tempo de treino, idioma e tamanho do corpus) por Liu e Lapata (2019), mantendo um foco comparativo com os resultados do modelo proposto na Subseção 5.1.

5.3 Estratégias de avaliação

Este estudo concentrou-se em realizar duas formas de avaliação dos resumos produzidos pelos modelos de sumarização: uma quantitativa com uso de métricas comuns da área e outra qualitativa com o auxílio de avaliadores humanos. Para a avaliação quantitativa dos resultados utilizamos a métrica ROUGE (Lin, 2004). Esses valores serviram como uma base para comparação entre os modelos e prepararam o terreno para a avaliação qualitativa.

Na avaliação qualitativa, a nossa proposta foi uma atividade de avaliação com pessoas por meio de um questionário. Esse questionário teve como objetivo capturar a percepção humana em relação aos textos gerados, permitindo uma análise mais profunda das qualidades dos resumos a partir da perspectiva dos avaliadores.

Nas Subseções seguintes ilustraremos essas etapas em detalhes.

5.3.1 Avaliação Quantitativa

Neste estudo, empregamos as métricas ROUGE-N (ROUGE-1 e ROUGE-2) e ROUGE-L para avaliar as pontuações de Recall e F1 entre os textos gerados e seus correspondentes textos-alvo. Realizamos essa avaliação para comparar o desempenho dos dois modelos treinados com o mesmo corpus em português. Como mencionado anteriormente, o primeiro modelo (baseado em LSTMs) serviu como um ponto de referência para comparação com o segundo modelo (baseado no *fine-tuning* do BERTimbau).

A métrica ROUGE-N tem como objetivo quantificar a quantidade de grupos de palavras correspondentes, chamados de n-gramas, entre o texto gerado pelo nosso modelo e uma referência. Aqui, um n-grama é uma sequência contínua de palavras. Para ilustrar, um unigrama (ou 1-grama) é uma única palavra, enquanto um bigrama (ou 2-grama) consiste em duas palavras consecutivas.

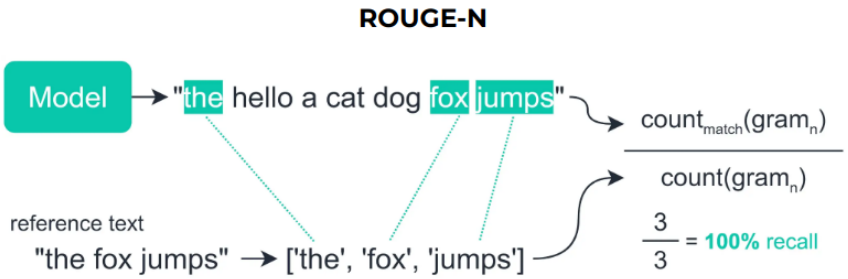


Figura 5.2: Exemplo de cálculo do *Recall* utilizando o Rouge-N (Briggs, 2021).

No contexto do ROUGE-N, o valor 'N' representa o número de palavras em um n-grama. Por exemplo, para o ROUGE-1, analisaremos a correspondência de unigramas entre o resultado gerado pelo modelo e a referência. Já para o ROUGE-2 e ROUGE-3, estaríamos avaliando correspondências de bigramas e trigramas, respectivamente. Além disso, uma vez que decidimos qual valor de 'N' utilizar, podemos calcular o *Recall* ou o *F1-Score*, ilustrado na Figura 5.2.

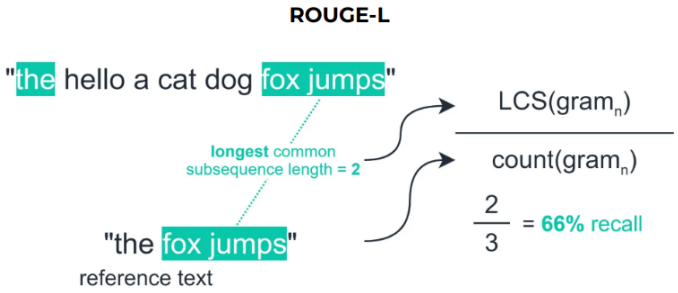


Figura 5.3: Exemplo de cálculo do *Recall* utilizando o Rouge-L (Briggs, 2021).

Em contrapartida, a métrica ROUGE-L concentra-se em medir a subsequência comum mais longa (LCS) entre a saída gerada pelo modelo e a referência. Basicamente, contamos o maior conjunto de palavras compartilhadas entre as duas sequências. A lógica por trás disso é que uma sequência compartilhada mais extensa pode indicar uma maior semelhança entre os dois textos. Podemos aplicar os mesmos cálculos de *Recall* e *F1-Score* como fizemos anteriormente, mas, desta vez, substituímos a correspondência pelo LCS na Figura 5.3.

5.3.2 Avaliação Qualitativa com Avaliadores

As métricas de avaliação quantitativas existentes podem não ser capazes de considerar completamente a semântica e a estrutura do texto. Mudanças na redação, com a introdução de palavras e frases novas, podem levar a um desempenho insatisfatório, uma vez que métodos automáticos podem não capturar

adequadamente o significado por completo. Portanto, para complementar à avaliação quantitativa, optamos por realizar uma avaliação com avaliadores humanos. Essa abordagem visa capturar de maneira mais apurada a qualidade dos resumos gerados, avaliando três aspectos: (i) o conteúdo propriamente dito; (ii) a estrutura lógica, que engloba elementos como objetivos, metodologia, resultados e conclusão; e, adicionalmente, proporcionar (iii) uma avaliação da percepção desses avaliadores sobre a autoria dos textos (se foram escritos por humanos ou não).

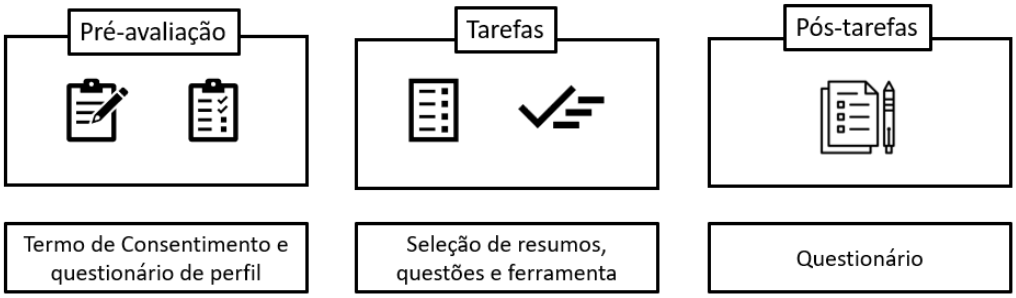


Figura 5.4: Etapas do procedimento para a avaliação dos textos gerados.

Para esta avaliação qualitativa, elaboramos o processo de avaliação ilustrado na Figura 5.4. O objetivo desse processo foi garantir que a avaliação fosse minuciosa e estivesse alinhada com as metas da pesquisa.

Sendo assim, na primeira etapa, de Pré-Avaliação, elaboramos e incorporamos um Termo de Consentimento ao questionário. Esse termo teve o propósito de assegurar a conformidade ética e a participação informada dos entrevistados. O Termo de Consentimento foi elaborado com a finalidade de fornecer aos participantes uma compreensão clara dos objetivos da pesquisa e esclarecer quaisquer riscos ou desconfortos potenciais que possam estar associados à sua participação. Além disso, enfatizamos a garantia de que os dados fornecidos pelos participantes seriam protegidos com rigor, assegurando sua anonimidade, privacidade e confidencialidade. Importante ressaltar que, para preservar a identidade individual, qualquer divulgação dos resultados ocorreria de forma agregada ou após o processo de análise.

Dentro do Termo de Consentimento, também enfatizamos a liberdade dos participantes em decidir participar voluntariamente, bem como a opção de recusar, interromper, desistir ou retirar o consentimento a qualquer momento, sem qualquer penalização. Somente após obter o consentimento informado, os participantes avançaram para a próxima etapa.

Ainda na etapa de Pré-avaliação, elaboramos o Questionário de Perfil do participante, solicitando dados adicionais, como nome e e-mail, embora

essa parte tenha sido opcional para preservar a confidencialidade. Além disso, coletamos outras informações como faixa etária e nível de escolaridade, para contextualizar as respostas fornecidas.

Com esses requisitos éticos e informativos atendidos, seguimos para a Etapa de Tarefas, em que selecionamos os resumos gerados e avaliamos as questões e a ferramenta de aplicação do questionário. A ferramenta utilizada foi o LimeSurvey[2], por nos permitir aleatorizar perguntas mais facilmente. Com isso, a estrutura da avaliação dos resumos ficou composta por cinco páginas sendo que cada página contém 7 perguntas avaliando 2 resumos, totalizando 35 perguntas e 10 resumos por participante. Vale ressaltar que nós reduzimos o número de perguntas para que o questionário não se tornasse muito extenso e trabalhoso com o intuito de que isso não interferisse na participação dos avaliadores. Ainda assim, o tempo médio de resposta do questionário foi de 24 minutos e 40 segundos.

Em mais detalhes, cada página de avaliação foi dividida em quatro subdivisões distintas:

1. **Comparação entre resumos:** Nessa subdivisão, os participantes tiveram a oportunidade de ler e analisar um par de resumos, conforme Figura A.1. Um dos resumos foi produzido por seres humanos ou uma IA, enquanto o outro foi gerado somente por IA.

2. **Avaliação do Primeiro Resumo:** Os participantes responderam a três perguntas específicas relacionadas ao primeiro resumo. Essas questões buscavam avaliar aspectos como a estrutura lógica do resumo (incluindo a presença de objetivos, metodologia e resultados), a compreensão do objetivo principal e a percepção da autoria humana do resumo.

3. **Avaliação do Segundo Resumo:** Similarmente à subdivisão anterior, três perguntas foram colocadas aos participantes com relação ao segundo resumo. Estas perguntas também exploraram a estrutura lógica, a compreensão do objetivo principal e a impressão sobre a autoria humana.

4. **Similaridade entre Resumos:** A última subdivisão continha uma única pergunta abordando a visão geral dos participantes sobre a similaridade de conteúdo entre os dois resumos analisados.

Ao entrar nas subdivisões Avaliação do Primeiro Resumo e Avaliação do Segundo Resumo, os participantes foram solicitados a responder sobre três diferentes aspectos:

[2]LimeSurvey: `https://www.limesurvey.org/pt-br`

a) Considero que o resumo tenha uma boa estrutura lógica (possui objetivo, metodologia e resultados)

b) Considero que eu consegui entender o objetivo principal do resumo

c) Considero que o resumo tenha sido elaborado por um ser humano.

A última parte do questionário consistiu na subdivisão 4, que teve uma abordagem mais abrangente. Nessa subdivisão, os participantes foram solicitados a expressar sua percepção geral sobre a similaridade de conteúdo entre os dois resumos que estavam analisando. Essa questão específica afirmava: "Em relação ao conteúdo, considero que os dois resumos acima apresentam semelhanças de conteúdo."A presença dessa subdivisão 4 em cada página do questionário é apresentada na Avaliação Geral, representada na Figura A.4.

As respostas fornecidas pelos participantes foram estruturadas de maneira objetiva, utilizando uma escala de avaliação conhecida como escala de Likert. Essa escala oferecia sete opções de resposta distintas, permitindo uma graduação das opiniões dos participantes com mais aspectos de neutralidade. A representação visual dessa escala, conforme exibida na Figura A.4, proporciona uma melhor visão da avaliação aplicado para cada resumo individualmente.

Os participantes foram organizados de maneira aleatória em quatro grupos distintos, cada um seguindo um caminho de avaliação específico. Cada grupo respondeu cinco páginas de perguntas, e cada página continha a análise de um par de resumos. Esses pares de resumos foram distribuídos conforme a seguinte lógica: os grupos 1, 3 e 4 avaliaram pares de resumos nos quais um deles foi produzido por um humano e o outro pela IA. Por outro lado, o grupo 2 avaliou dois resumos diferentes, ambos gerados pela IA a partir do mesmo texto original.

O grupo 2 foi uma criação intencional para explorar um fenômeno interessante. Mesmo após serem informados de que ambos os resumos poderiam ser gerados pela IA, os participantes, ainda durante a prévia do questionário, tendiam a associar a autoria de um dos resumos a um ser humano. Isso permitiu examinar a percepção dos participantes e entender melhor como suas expectativas e preconcepções podem influenciar a interpretação da autoria dos resumos.

Por fim, o questionário contou com uma caixa de texto em aberto para que os participantes registrassem de forma opcional seus *feedbacks*, sugestões e percepções sobre os resumos e sobre a avaliação como um todo. Os detalhes de todas as subdivisões estabelecidas, assim como Termo de Consentimento e perfil do participante no questionário estão disponíveis no Apêndice A.

6
Resultados

Neste Capítulo, apresentamos os resultados obtidos tanto da avaliação quantitativa quando da avaliação qualitativa dos dois modelos de sumarização implementados. Os resultados da avaliação quantitativa está descrito na Seção 6.1 e apresenta a diferença de desempenho entre os dois modelos utilizando o ROUGE. Enquanto a Seção 6.2 traz os resultados da avaliação dos participantes sobre os resumos do melhor modelo.

Ambas propostas de modelos foram replicadas, mas utilizando o nosso corpus em português como insumo de treino e teste. Os modelos receberam como entrada os pares: corpo e resumo de cada artigo do corpus criado. Embora a estrutura geral dos modelos tenha permanecido inalterada, ao longo dos experimentos, fizemos ajustes sutis de hiper-parâmetros a fim de melhorar os resultados. Um desses ajustes foi equalizar o tamanho do vocabulário em 30 mil tokens e do tamanho máximo de tokens de entrada em até 3 mil tokens. Outros ajustes foram feitos e serão apresentados ao longo desse Capítulo.

Também limitamos o tempo de treinamento para ambos os modelos, para garantir maior paridade na comparação entre eles. Os treinamentos foram realizados utilizando o Cluster HPC ExACTa PUC-Rio, iniciativa ExACTa que oferece um ambiente para acesso único de recursos computacionais (infraestrutura hiperconvergente, nós de computação com GPUs, gerenciador de recursos com filas de trabalhos e acesso via web). Limitamos o tempo de treino a 7 dias, usando uma placa gráfica Titan RTX 3090 da Nvidia com 130 Tensor TFLOPs de desempenho, 576 tensor cores e 24GB de memória GDDR6[1].

6.1
Desempenho dos modelos

Como esperado, o modelo de sumarização com *fine-tuning* do BERT apresentou um desempenho notavelmente melhor em relação ao modelo baseado em LSTM. Os valores de *Recall* (ROUGE-1, ROUGE-2 e ROUGE-L) foram substancialmente altos, com pontuações de 59,80, 45,15 e 54,92, respectivamente. Esses resultados superaram significativamente as pontuações de *Recall* do modelo LSTM, que foram de 25,74, 3,96 e 23,19, respectivamente. Além disso, ao

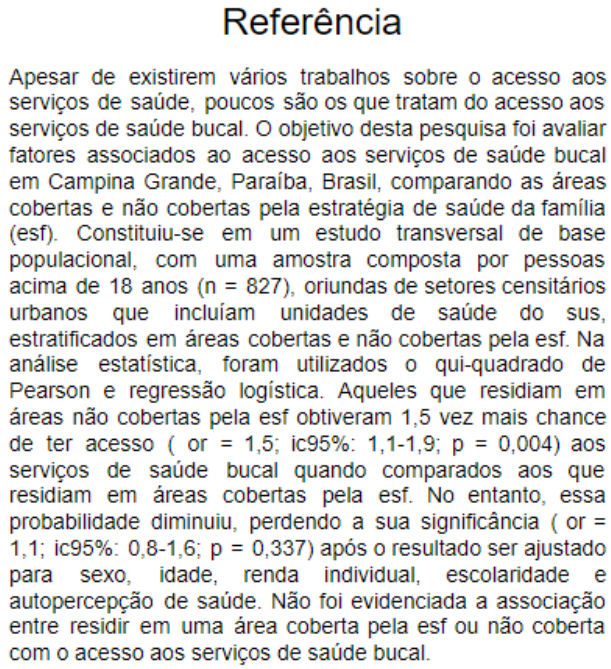
[1] `https://www.nvidia.com/pt-br/deep-learning-ai/products/titan-rtx/`

considerar o *F1-score*, o modelo baseado no BERT também se destacou significativamente, indicando uma melhoria geral em relação ao modelo utilizado como *baseline*.

Tabela 6.1: ROUGE-1, ROUGE-2 e ROUGE-L Recall e F1-Score como resultados no nosso conjunto de teste em português.

Modelo	**Recall**			**F1-Score**		
	R1	**R2**	**RL**	**R1**	**R2**	**RL**
LSTM (Cohan et al., 2018)	25.74	03.96	23.19	29.59	04.59	26.59
BERT (Liu e Lapata, 2019)	59.80	45.15	54.92	65.17	49.11	58.87

Com esses resultados, observamos também que apenas o modelo baseado no BERTimbau foi capaz de gerar resumos que incorporam informações cruciais dos artigos de entrada, como as citações numéricas. Como por exemplo, a citação de tamanho de amostra, mês e ano do estudo, idades de pacientes e em alguns casos o intervalo de confiança do estudo. A Figura 6.2 traz um resumo gerado pelo modelo baseado no BERTimbau em que há faixa etária e faixa de renda dos participantes. Isso ressalta a habilidade do modelo em detectar e incorporar parte dos dados numéricos relevantes de um artigo científico mesmo sob as condições impostas nesse estudo de limitação de tempo, hardware e tamanho do corpus. Essa característica é de especial interesse no contexto científico, onde informações quantitativas desempenham um papel fundamental na apresentação dos experimentos e resultados.


Referência

Apesar de existirem vários trabalhos sobre o acesso aos serviços de saúde, poucos são os que tratam do acesso aos serviços de saúde bucal. O objetivo desta pesquisa foi avaliar fatores associados ao acesso aos serviços de saúde bucal em Campina Grande, Paraíba, Brasil, comparando as áreas cobertas e não cobertas pela estratégia de saúde da família (esf). Constituiu-se em um estudo transversal de base populacional, com uma amostra composta por pessoas acima de 18 anos (n = 827), oriundas de setores censitários urbanos que incluíam unidades de saúde do sus, estratificados em áreas cobertas e não cobertas pela esf. Na análise estatística, foram utilizados o qui-quadrado de Pearson e regressão logística. Aqueles que residiam em áreas não cobertas pela esf obtiveram 1,5 vez mais chance de ter acesso (or = 1,5; ic95%: 1,1-1,9; p = 0,004) aos serviços de saúde bucal quando comparados aos que residiam em áreas cobertas pela esf. No entanto, essa probabilidade diminuiu, perdendo a sua significância (or = 1,1; ic95%: 0,8-1,6; p = 0,337) após o resultado ser ajustado para sexo, idade, renda individual, escolaridade e autopercepção de saúde. Não foi evidenciada a associação entre residir em uma área coberta pela esf ou não coberta com o acesso aos serviços de saúde bucal.

Figura 6.1: Referência para os resumos gerados da Figura 6.2.

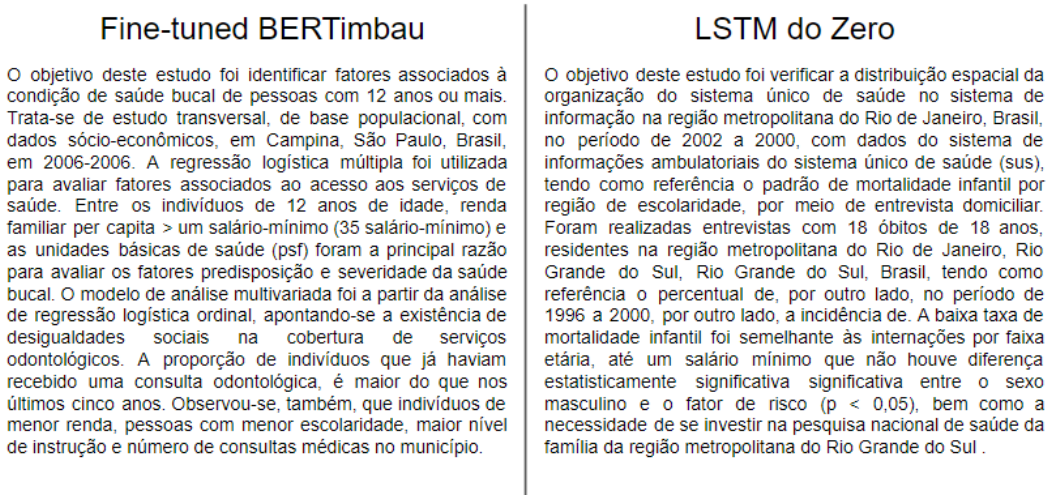


Figura 6.2: Exemplo 01 dos resumos gerados pelos dois modelos.

Outra característica importante a mencionar é que, ao compararmos o conteúdo do resumo gerado pelo BERTimbau com o resumo de referência na Figura 6.1, podemos observar que ambos abordam a condição de saúde bucal. Isso sugere uma semelhança entre os conteúdos e, portanto, indica uma alta qualidade informativa do que foi produzido.

Apesar dos resultados promissores obtidos pelo modelo baseado no BERTimbau, é importante ressaltar que algumas inconsistências foram identificadas durante a análise das predições dos dois modelos. Ocasionalmente, os modelos geraram palavras repetitivas, identificaram erroneamente a localização do estudo e produziram valores numéricos distorcidos em relação aos resumos de referência. Por exemplo, na Figura 6.2, o período informado do estudo vai de 2006 a 2006, enquanto o resumo de referência não menciona o intervalo. Outra ressalva é quanto ao local do estudo, o município "Campina"foi erroneamente interpretado como parte do estado de São Paulo, quando na verdade o estudo foi originalmente conduzido no município de Campina Grande, no estado da Paraíba.

Essas diferenças podem causar confusão na busca pelo conteúdo resumido. No entanto, é importante destacar que, em geral, o resumo mantém uma coerência textual, apresentando informações sobre os objetivos, resultados e conclusões conforme observado pelos autores.

Na avaliação qualitativa, notamos que 71,43% das pessoas que avaliaram esse resumo concordam que ele poderia ter sido redigido por um ser humano (a maior porcentagem entre os resumos avaliados). Adicionalmente, 92,86% dessas pessoas estão de acordo que este resumo, gerado pelo nosso modelo, apresenta uma estrutura lógica sólida e que conseguiram compreender seu objetivo principal.

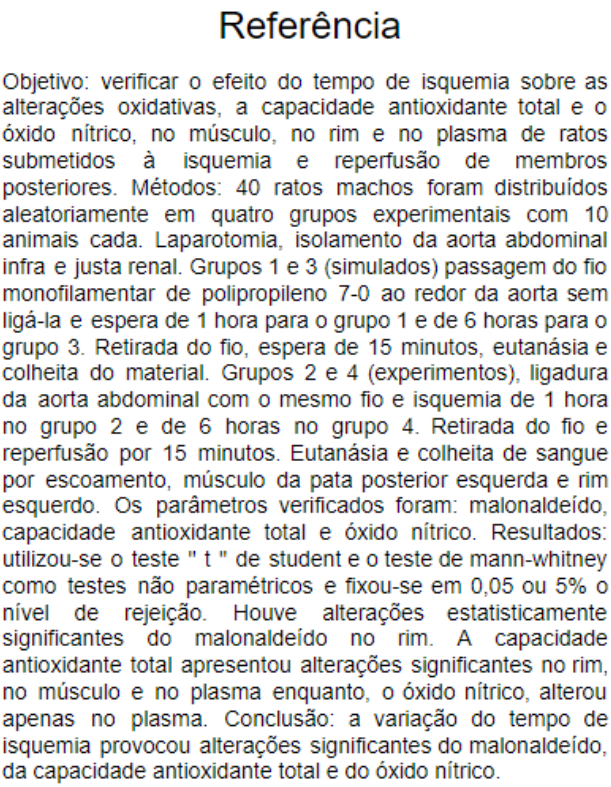


Figura 6.3: Referência para os resumos gerados da Figura 6.4.

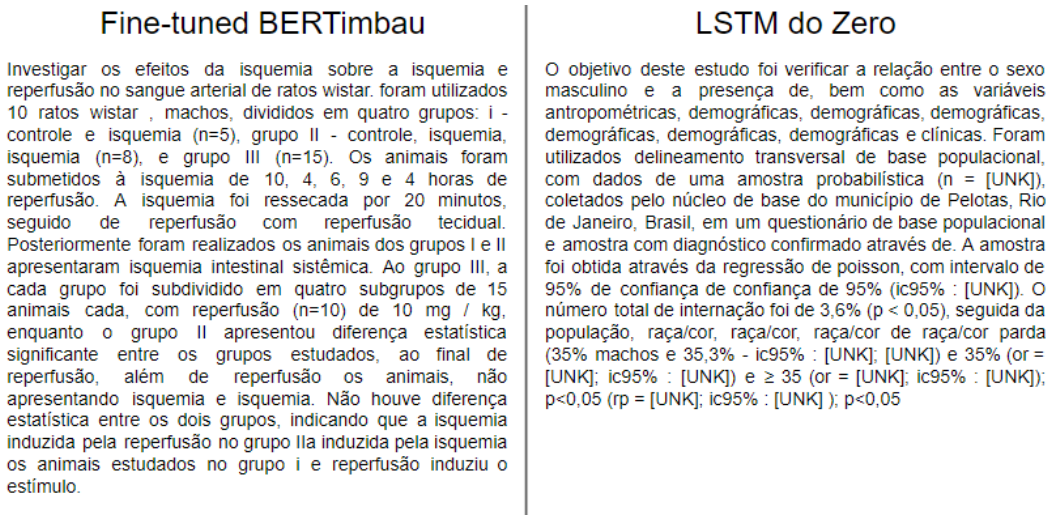


Figura 6.4: Exemplo 02 dos resumos gerados pelos dois modelos.

A Figura 6.4 apresenta o segundo exemplo de previsão gerado pelos modelos. É evidente que este texto exibe falta de coerência em alguns trechos. Por exemplo, ao mencionar os quatro grupos experimentais, o modelo apresenta três deles, sem explicar qual seria o quarto grupo. Além disso, há uma medição incorreta de 15 ratos quando anteriormente o resumo informa que apenas 10 foram utilizados em todo o experimento, juntamente com outras falhas na continuidade do texto e repetições.

Apesar disso, semelhante aos exemplos anteriores, este exemplo apresenta em sua estrutura informações sobre o objetivo e a metodologia do estudo. No entanto, ao compararmos este resumo com a referência na Figura 6.3, os resultados e a conclusão são ofuscados pelas inconsistências que, de certa forma, dificultam a compreensão desses pontos. Essas modificações podem refletir a ausência de artigos sobre esse tema ou a falta dessa nomenclatura no corpus utilizado pelo modelo.

A Figura 6.4 também evidencia outros exemplos de inconsistências nos resumos produzidos pelo LSTM, que incluem tokens não identificados pelo modelo (representados por "*[UNK]*"). Essa representação mostra situações em que o modelo não conseguiu prever as palavras seguintes.

Entre os avaliadores deste resumo, não houve acordo quanto à sua autoria, sendo que metade acredita que tenha sido feito por uma pessoa. Ao mesmo tempo, 50% concordam que ele possui uma estrutura lógica adequada e entenderam claramente o seu propósito principal.

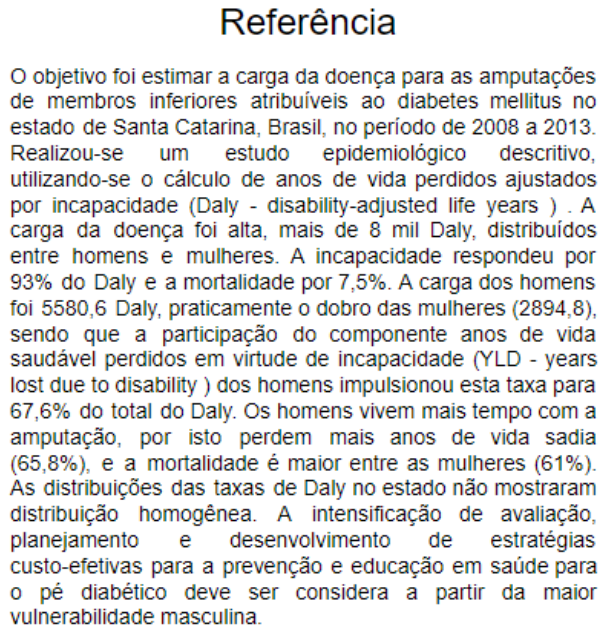


Figura 6.5: Referência para os resumos gerados da Figura 6.6.

Na nossa perspectiva, o resumo gerado pelo modelo baseado no BERTimbau, conforme ilustrado na Figura 6.5, apresenta elementos de estrutura lógica, como objetivos, metodologia e resultados. No entanto, esses elementos não estão devidamente explicados. Por exemplo, o objetivo mencionado na primeira sentença não esclarece qual doença específica está sendo estudada. Além disso, os resultados são prejudicados pela ausência de citações numéricas, como a taxa de amputação em pacientes com diabetes, a diferença de anos entre homens e a frequência de amputações em mulheres.

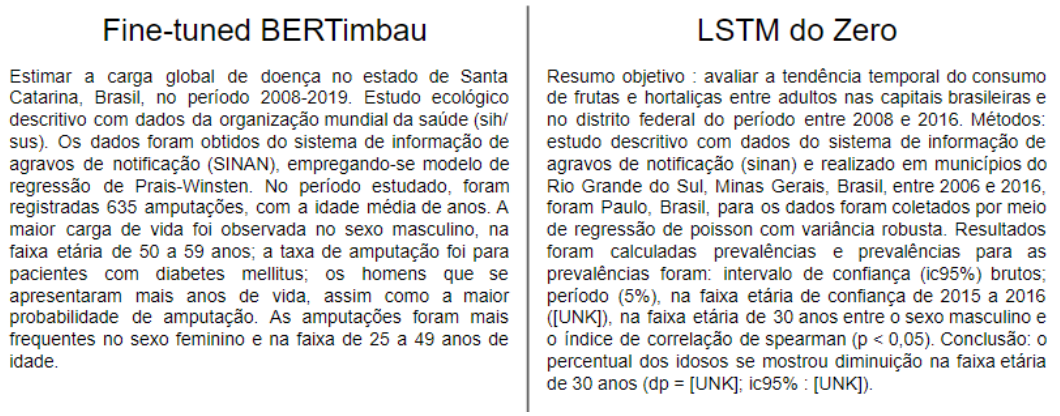


Figura 6.6: Exemplo 03 dos resumos gerados pelos dois modelos.

Apesar dessas limitações, entre os 11 participantes que avaliaram esse resumo, 9 deles (ou 90,91%) concordaram que a estrutura lógica do resumo é adequada, e 81,82% afirmaram que entenderam seu objetivo principal. É interessante notar que 81,8% dos participantes consideraram que o conteúdo é similar ao resumo de referência na Figura 6.5. No entanto, apenas 18,18% (ou seja, 2 pessoas) acreditaram que esse resumo foi elaborado por um ser humano. Essa é a taxa de autoria mais baixa entre todos os resumos avaliados no questionário.

De maneira geral, as métricas usadas para avaliar os dois modelos que foram propostos mostram que o modelo que passou por um processo de *fine-tuning* teve um desempenho superior ao modelo LSTM quando submetidos ao mesmo tipo de treinamento. Com base nisso, decidimos avaliar qualitativamente apenas os resumos gerados pelo modelo BERTimbau. Essa avaliação foi feita por meio do questionário, conforme descrito na Seção 5.3.2, com mais detalhes disponíveis no Apêndice A.

6.2 Avaliação qualitativa

Nosso estudo também incluiu a realização de uma avaliação qualitativa dos resumos produzidos pelo modelo baseado no BERTimbau. Essa avaliação envolveu a análise de dois tipos de resumos: o primeiro, sendo o resumo proveniente do artigo científico de referência, e o segundo, um resumo gerado pela IA. Os resultados dessa avaliação estão detalhados nas próximas seções.

6.2.1 Perfil dos participantes

Um total de 36 indivíduos participou na avaliação dos resumos gerados pelo modelo baseado no BERTimbau. A divisão por faixas etárias resultou nas

seguintes categorias: até 17 anos, 18 a 24 anos, 25 a 35 anos, 36 a 50 anos e 51 anos ou mais. A maioria dos participantes se enquadram na faixa etária de 25 a 35 anos, compreendendo um total de 32 pessoas (88,89% do total). Enquanto isso, três pessoas (8,33%) estão entre os 36 e 50 anos, e somente uma pessoa (2,78%) tem mais de 50 anos. Essa discrepância nos leva a concluir que as respostas estratificadas por idade tem pouca relevância para essa avaliação.

No que diz respeito ao nível educacional, observa-se uma distribuição mais uniforme entre as categorias, que foram delineadas da seguinte forma: sem escolaridade (sem respostas), fundamental incompleto (sem respostas), fundamental completo (sem respostas), ensino médio incompleto (sem respostas), ensino médio completo (1 resposta), ensino superior incompleto (2 respostas), ensino superior completo (16 respostas), mestrado incompleto (4 respostas), mestrado completo (6 respostas), doutorado incompleto (5 respostas) e doutorado completo (2 respostas), conforme ilustrado na Figura 6.7.

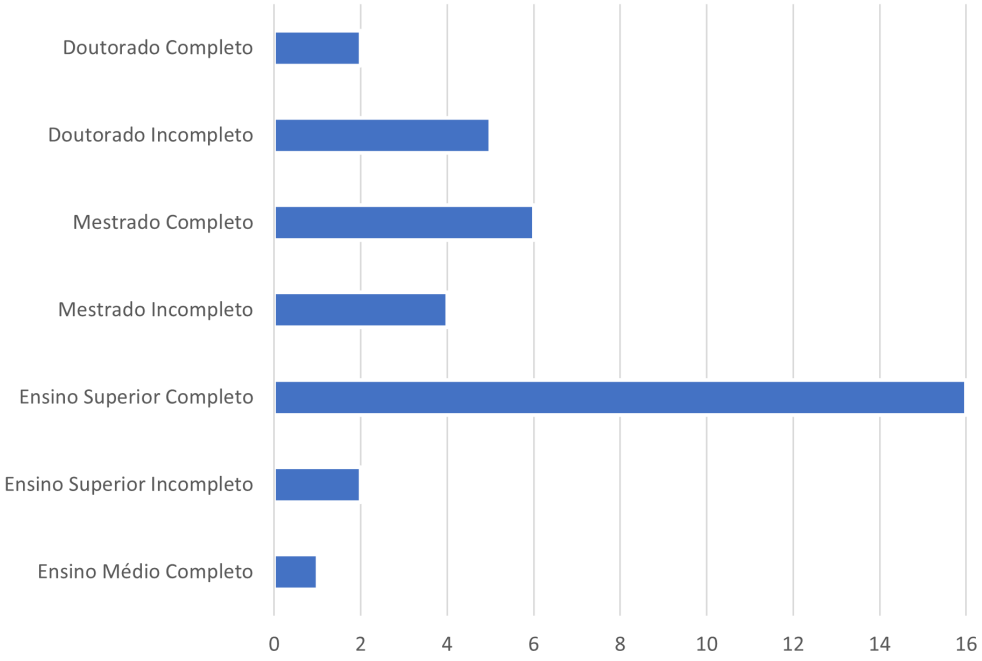


Figura 6.7: Distribuição dos participantes por grau de escolaridade.

6.2.2 Análise das respostas

A análise da Figura 6.8 reflete a média das percepções coletadas dos 36 participantes que responderam ao questionário. Essa figura indica que, de maneira abrangente, os resumos gerados pela IA são geralmente considerados como apresentando uma estrutura lógica coesa bem como os resumos produzidos por seres humanos.

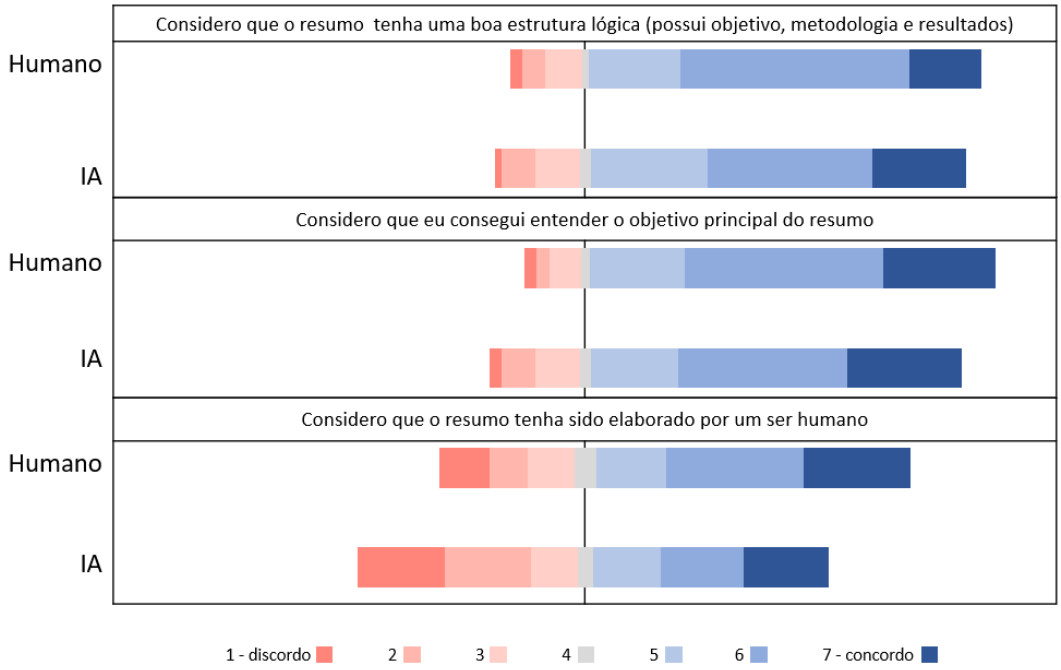


Figura 6.8: Média de todas as 36 respostas sobre os resumos elaborados tanto por humanos quanto pela IA em todos os grupos (1, 2, 3 e 4).

Adicionalmente, é interessante observar que a avaliação global relacionada ao objetivo principal de cada resumo demonstra que os resumos elaborados por indivíduos humanos são ligeiramente mais percebidos que os resumos gerados pela IA. Isso sugere que, em termos de transmitir a mensagem central, ambos os conjuntos de resumos são comparáveis em termos de mensagem percebida.

Outra dimensão interessante da avaliação recai sobre a percepção dos participantes acerca da autoria dos resumos. O questionário revela que, na maioria das instâncias, os participantes tendem a acreditar que os resumos foram criados por seres humanos, mesmo quando se trata dos resumos produzidos pelo modelo que se baseia no BERTimbau. Esse resultado sugere que a qualidade de escrita dos resumos gerados pela IA pode ser alta, ao ponto de suscitar uma confusão perceptível entre os participantes, que não conseguem diferenciar entre as origens humanas e automatizadas dos resumos.

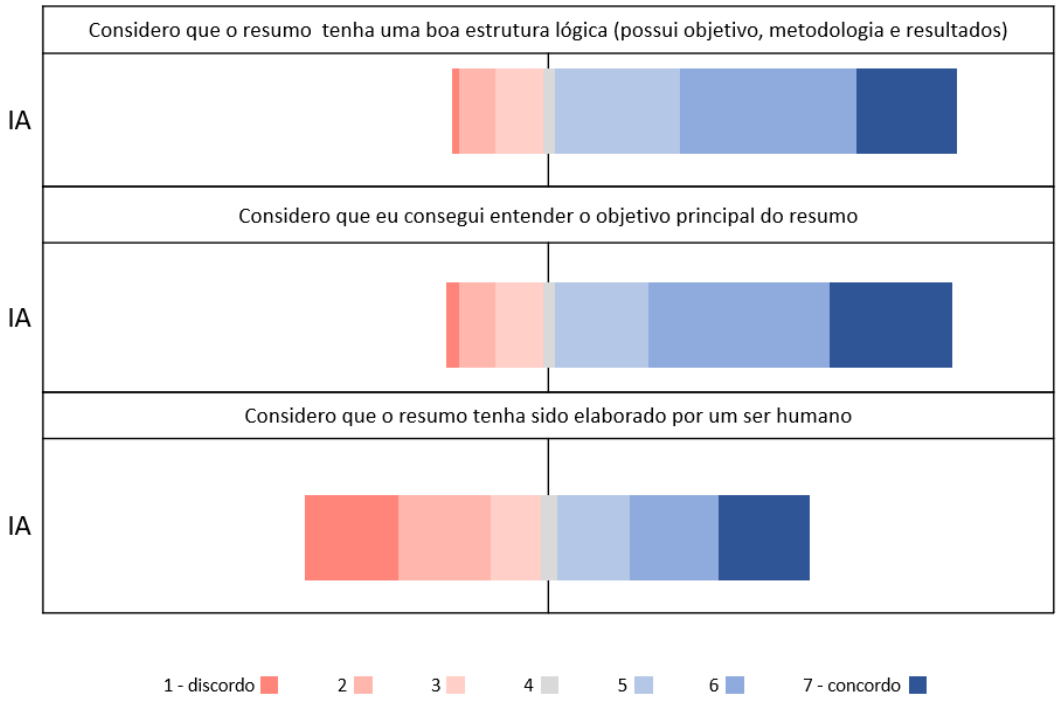


Figura 6.9: Média das respostas sobre os resumos do grupo 2 (elaborados apenas pela IA).

Ao focarmos exclusivamente nas respostas provenientes do Grupo 2, o qual abrange apenas os resumos gerados de forma automática, uma análise mais minuciosa da Figura 6.9 permite perceber que a estrutura lógica dos resumos demonstrou bom grau de clareza na maioria dos casos. Além disso, é notável que, em grande parte das situações, o objetivo principal de cada resumo foi bem compreendido.

É relevante notar que, apesar dessas nuances, um outro aspecto emerge a partir das avaliações: os participantes associaram os resumos tanto a autoria humana quanto automática. Essa percepção é intrigante, considerando-se que os resumos em questão foram gerados automaticamente pelo modelo baseado no BERTimbau e os participantes foram alertados aos fato de que os dois resumos poderiam ser da IA.

Por fim, procedemos à avaliação da percepção dos participantes no que diz respeito à semelhança de conteúdo entre os resumos, conforme ilustrado com a Figura 6.10. Ao considerarmos a análise da Avaliação Geral, abrangendo as respostas de todos os grupos, constatamos que, em sua maioria, os conteúdos presentes nos dois tipos de resumos apresentam graus de similaridade consideráveis. Isso implica que o modelo demonstrou um bom desempenho ao reproduzir adequadamente o conteúdo abordado nos artigos científicos originais, quando confrontado com os resumos produzidos por seus respectivos autores.

Uma investigação mais específica das respostas agregadas provenientes

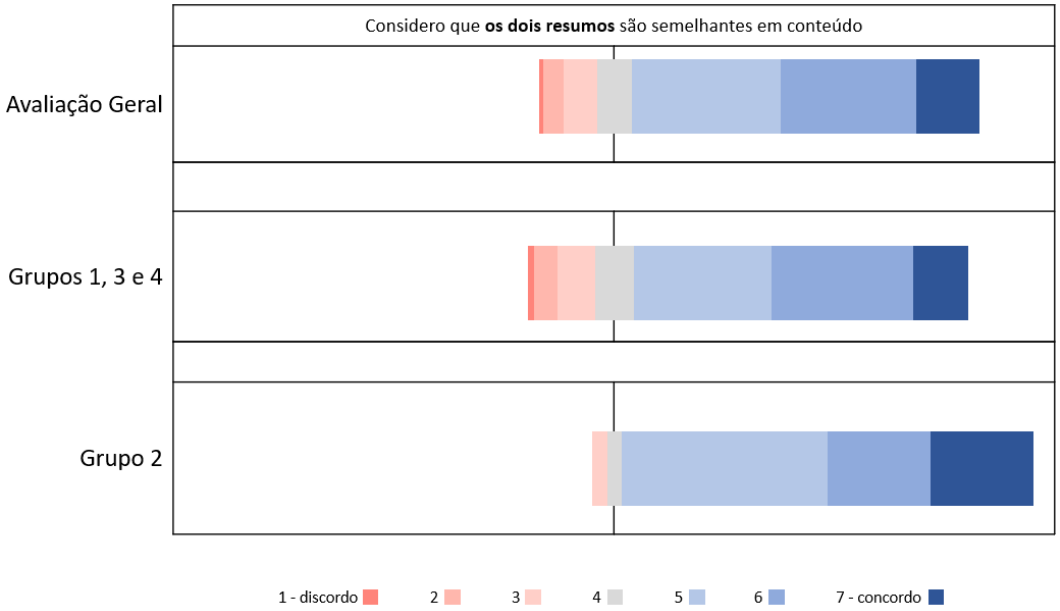


Figura 6.10: Média das respostas sobre a semelhança entre os conteúdos apresentados pelos dois resumos.

dos grupos 1, 3 e 4, que englobam tanto resumos produzidos por seres humanos quanto pela IA, evidencia uma continuidade em relação à tendência observada na média global. Em outras palavras, a tendência de similaridade de conteúdo se mantém consistentemente próxima à média geral, indicando que o modelo não apenas reproduziu a estrutura lógica dos textos e o conteúdo de maneira eficaz, mas também manteve essa capacidade de maneira consistente ao considerarmos diferentes fontes de resumos.

7
Conclusões

Neste estudo, realizamos uma análise comparativa entre dois modelos distintos para a criação de resumos abstrativos em língua portuguesa, buscando gerá-los automaticamente a partir de textos técnicos das Ciências da Saúde (nossa questão de pesquisa principal). Para isso, buscamos na literatura pelas técnicas mais recentes de sumarização de textos longos e escolhemos duas (buscando responder a QP1): um modelo é fundamentado em redes neurais LSTM (Cohan et al., 2018), enquanto o outro se baseia no conceito de *fine-tuning* de um LLM (Liu e Lapata, 2019). Nossa investigação teve como alvo principal documentos científicos, que serviram como o alicerce do nosso corpus. Através deste trabalho, conseguimos demonstrar a eficácia do método de *fine-tuning* de um modelo pré-treinado para o português brasileiro, que superou o desempenho do modelo LSTM treinado *from the scratch*. Vale destacar que nossos resultados ratificam a literatura já existente, porém representando os primeiros resultados registrados que mostram ser possível resumir textos longos em língua portuguesa (viabilizando nossa QP2).

Com o propósito de viabilizar nossos experimentos, uma vez que não encontramos corpora para sumarização de textos longos em português, desenvolvemos um *web crawler* que nos permitiu coletar um total de 7.450 artigos científicos da base aberta da SciElo-BR. Esse corpus juntamente com o código-fonte do *web crawler* encontram-se à disposição da comunidade científica para que os resultados possam ser replicados e investigados de maneira mais profunda. Adicionalmente, é possível utilizar nosso *web crawler* para expandir o conjunto de dados obtidos do SciElo-BR, incluindo artigos de áreas além das Ciências da Saúde, tornando o corpus ainda mais abrangente e diversificado.

No que tange aos modelos treinados, os resultados obtidos com base nas métricas ROUGE, a partir do processo de *fine-tuning* do modelo BERT para a língua portuguesa, são notáveis e promissores, principalmente considerando o esforço relativamente baixo exigido (corpus menor em comparação aos existentes em língua inglesa e limitação no tempo de treino dos modelos). Isso sublinha a razão por trás da evolução dos modelos de sumarização, que evoluíram de Redes Neurais Recorrentes (RNNs) tradicionais para Long Short-Term Memory (LSTM) e agora se apoiam em LLMs, como o BERT (Devlin

et al., 2019). É relevante destacar que os modelos baseados em RNNs ou LSTMs frequentemente requerem uma grande quantidade de dados de treinamento e esforços intensivos de tempo e hardware nesse processo. Esse requisito torna-se ainda mais evidente em cenários com poucos exemplos disponíveis, como é o caso do nosso estudo. Em contrapartida, o modelo BERTimbau demonstrou sua eficácia mesmo com um corpus menor e um processo de *fine-tuning*, consolidando sua habilidade em resolver tarefas de PLN em língua portuguesa e, nesse caso, produzindo resumos.

Embora o ROUGE tenha se mostrado um método eficaz para avaliação automática, nossa pesquisa não se limitou à geração de resumos, mas também submeteu esses resumos a análises e avaliações por parte de seres humanos. Essa abordagem permitiu a validação da capacidade do modelo de gerar textos bem estruturados e úteis, mesmo com um corpus modesto em comparação a outros da língua inglesa, junto com recursos de hardware limitados para treinamento, validação e testes. Esses resumos não apenas auxiliam na filtragem de conteúdo relevante em meio a um grande volume de dados não estruturados, mas também são capazes de fazê-lo mesmo após a utilização de um mecanismo de busca na web.

Dessa forma, ao analisar as percepções obtidas através das respostas ao questionário, obtivemos uma compreensão mais profunda de como os resumos produzidos por diferentes fontes são percebidos (resolvendo assim a QP3). Além disso, destacamos como a qualidade da escrita do modelo baseado no BERTimbau pode ser tão boa a ponto de ser comparada, em termos de autoria percebida, aos resumos produzidos por seres humanos. Essas observações também evidenciam como as redes neurais profundas podem aprender os padrões de escrita a ponto de reproduzirem a estrutura lógica dos resumos de documentos científicos, incluindo objetivos, metodologia e resultados, conferindo aos textos gerados um toque adicional de "humanização".

Além da qualidade percebida pelos participantes, esses resumos carregam o conteúdo essencial dos textos longos fornecidos como entrada para o modelo. Quando comparamos os resumos gerados com o texto de referência, a percepção geral indicou que os modelos compartilhavam conteúdo similar, o que destaca mais um ponto positivo para a nossa geração automática, ou seja, isso contribui ainda mais para a percepção de que esses resumos poderiam ter sido criados por seres humanos. Dessa forma, demonstramos como resumir automaticamente documentos técnicos de Ciências da Saúde em língua portuguesa.

Em termos de trabalhos futuros, temos a intenção de conduzir avaliações qualitativas dos resumos gerados, contando com a participação de especialistas no campo. Esses especialistas podem fornecer *insights* valiosos sobre a presença

ou ausência de terminologia técnica e como isso pode ter afetado o conteúdo dos resumos. Além disso, pretendemos compreender o desempenho do modelo quando aplicado a outros tipos de documentos técnicos dentro do mesmo domínio das Ciências da Saúde, como relatórios ou manuais médicos. Essa avaliação, essencial para o desenvolvimento contínuo deste trabalho, precisa ser realizada dentro de um contexto interdisciplinar para garantir uma abordagem abrangente e eficaz.

No que diz respeito aos próximos passos, pretendemos explorar a capacidade de resumir textos utilizando outros modelos multilíngues de grande porte, como o GPT (Radford et al., 2018) e o BARD (Thoppilan et al., 2022). Essa abordagem ainda exigiria um esforço mínimo, ao mesmo tempo em que poderia potencialmente resultar em um desempenho aprimorado com o propósito de atualizar e aperfeiçoar as habilidades e a eficácia do nosso modelo de sumarização no futuro.

Referências Bibliográficas

Aleixo, P. e Pardo, T. A. S. (2008). Cstnews: um córpus de textos jornalísticos anotados segundo a teoria discursiva multidocumento cst (cross-document structure theory.

An, C., Zhong, M., Chen, Y., Wang, D., Qiu, X., e Huang, X. (2021). Enhancing scientific papers summarization with citation graph. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 12498–12506.

Briggs, J. (2021). Measure nlp accuracy with rouge | towards data science. `https://towardsdatascience.com/the-ultimate-performance-metric-in-nlp-111df6c64460`. (Accessed on 12/04/2022).

Callison-Burch, C., Osborne, M., e Koehn, P. (2006). Re-evaluating the role of bleu in machine translation research. In *11th conference of the european chapter of the association for computational linguistics*, pages 249–256.

Carmo, D., Piau, M., Campiotti, I., Nogueira, R., e Lotufo, R. (2020). Ptt5: Pretraining and validating the t5 model on brazilian portuguese data. *arXiv preprint arXiv:2008.09144*.

Cheng, F. e Miyao, Y. (2017). Classifying temporal relations by bidirectional lstm over dependency paths. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1–6.

Chiu, J. P. e Nichols, E. (2016). Named entity recognition with bidirectional lstm-cnns. *Transactions of the association for computational linguistics*, 4:357–370.

Clark, K., Luong, M.-T., Le, Q. V., e Manning, C. D. (2020). Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*.

Cohan, A., Dernoncourt, F., Kim, D. S., Bui, T., Kim, S., Chang, W., e Goharian, N. (2018). A discourse-aware attention model for abstractive summarization of long documents. *arXiv preprint arXiv:1804.05685*.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., e Stoyanov, V. (2019). Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

de Vargas Feijó, D. e Moreira, V. P. (2018). Rulingbr: A summarization dataset for legal texts. In *Computational Processing of the Portuguese Language: 13th International Conference, PROPOR 2018, Canela, Brazil, September 24–26, 2018, Proceedings 13*, pages 255–264. Springer.

Devlin, J., Chang, M.-W., Lee, K., e Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Durrett, G., Berg-Kirkpatrick, T., e Klein, D. (2016). Learning-based single-document summarization with compression and anaphoricity constraints. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1998–2008, Berlin, Germany. Association for Computational Linguistics.

Else, H. (2020). How a torrent of COVID science changed research publishing — in seven charts. *Nature*, 588(7839):553–553.

Fabbri, A. R., Kryściński, W., McCann, B., Xiong, C., Socher, R., e Radev, D. (2021). SummEval: Re-evaluating Summarization Evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

Fabbri, A. R., Li, I., She, T., Li, S., e Radev, D. R. (2019). Multi-news: A large-scale multi-document summarization dataset and abstractive hierarchical model. *arXiv preprint arXiv:1906.01749*.

Graff, D., Kong, J., Chen, K., e Maeda, K. (2003). English gigaword. *Linguistic Data Consortium, Philadelphia*, 4(1):34.

Grusky, M., Naaman, M., e Artzi, Y. (2018). Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. *arXiv preprint arXiv:1804.11283*.

Gupta, S. e Gupta, S. K. (2019). Abstractive summarization: An overview of the state of the art. *Expert Systems with Applications*, 121:49–65.

Hasan, T., Bhattacharjee, A., Islam, M. S., Mubasshir, K., Li, Y.-F., Kang, Y.-B., Rahman, M. S., e Shahriyar, R. (2021). Xl-sum: Large-scale multilingual abstractive summarization for 44 languages. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4693–4703.

Hermann, K. M., Kocisky, T., Grefenstette, E., Espeholt, L., Kay, W., Suleyman, M., e Blunsom, P. (2015). Teaching machines to read and comprehend. *Advances in neural information processing systems*, 28.

Ibrahim Altmami, N. e El Bachir Menai, M. (2022). Automatic summarization of scientific articles: A survey. *Journal of King Saud University - Computer and Information Sciences*, 34(4):1011–1028.

Inácio, M. L. (2021). *Sumarização de Opinião com base em Abstract Meaning Representation*. PhD thesis, Universidade de São Paulo.

Kalinowski, M., Escovedo, T., Villamizar, H., e Lopes, H. (2023). *Engenharia de Software para Ciência de Dados: Um guia de boas práticas com ênfase na construção de sistemas de Machine Learning em Python*. Casa do Código.

Kim, B., Kim, H., e Kim, G. (2018). Abstractive summarization of reddit posts with multi-level memory networks. *arXiv preprint arXiv:1811.00783*.

Kim, J.-H., Lee, S.-W., Kwak, D., Heo, M.-O., Kim, J., Ha, J.-W., e Zhang, B.-T. (2016). Multimodal residual learning for visual qa. *Advances in neural information processing systems*, 29.

Kornilova, A. e Eidelman, V. (2019). Billsum: A corpus for automatic summarization of us legislation. *arXiv preprint arXiv:1910.00523*.

Krizhevsky, A., Sutskever, I., e Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25.

Ladhak, F., Durmus, E., Cardie, C., e McKeown, K. (2020). WikiLingua: A new benchmark dataset for cross-lingual abstractive summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4034–4048, Online. Association for Computational Linguistics.

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., e Soricut, R. (2019). Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Lin, C.-Y. (2004). Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Liu, Y. (2019). Fine-tune bert for extractive summarization. *arXiv preprint arXiv:1903.10318*.

Liu, Y. e Lapata, M. (2019). Text summarization with pretrained encoders. *CoRR*, abs/1908.08345.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., e Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Lloret, E., Romá-Ferri, M. T., e Palomar, M. (2013). Compendium: A text summarization system for generating abstracts of research papers. *Data & Knowledge Engineering*, 88:164–175.

Ma, T., Pan, Q., Rong, H., Qian, Y., Tian, Y., e Al-Nabhan, N. (2021). T-bertsum: Topic-aware text summarization based on bert. *IEEE Transactions on Computational Social Systems*, 9(3):879–890.

Martins, A. F. e Smith, N. A. (2009). Summarization with a joint model for sentence extraction and compression. In *Proceedings of the Workshop on Integer Linear Programming for Natural Language Processing*, pages 1–9.

Melis, G., Dyer, C., e Blunsom, P. (2017). On the state of the art of evaluation in neural language models. *CoRR*, abs/1707.05589.

Mittal, A. (2019). Understanding rnn and lstm. what is neural network? `https://aditi-mittal.medium.com/understanding-rnn-and-lstm-f7cdf6dfc14e`. (Accessed on 08/23/2023).

Narayan, S., Cohen, S. B., e Lapata, M. (2018). Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. *arXiv preprint arXiv:1808.08745*.

Nóbrega, F. A. A. e Pardo, T. A. S. (2016). Investigating machine learning approaches for sentence compression in different application contexts for portuguese. In *International Conference on Computational Processing of the Portuguese Language*, pages 245–250. Springer.

Noriega, L. (2005). Multilayer perceptron tutorial. *School of Computing. Staffordshire University*, 4(5):444.

Paiola, P. H., de Rosa, G. H., e Papa, J. P. (2022). Deep learning-based abstractive summarization for brazilian portuguese texts. In *Intelligent Systems: 11th Brazilian Conference, BRACIS 2022, Campinas, Brazil, November 28–December 1, 2022, Proceedings, Part II*, pages 479–493. Springer.

Papineni, K., Roukos, S., Ward, T., e Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., et al. (2018). Improving language understanding by generative pre-training.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P. J., et al. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67.

Raschka, S. (2018). Mlxtend: Providing machine learning and data science utilities and extensions to python's scientific computing stack. *The Journal of Open Source Software*, 3(24).

Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386.

Saggion, H. e Lapalme, G. (2000). Selective analysis for automatic abstracting: Evaluating indicativeness and acceptability. In *RIAO*, pages 747–764. Citeseer.

Sanh, V., Debut, L., Chaumond, J., e Wolf, T. (2019). Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Sharma, E., Li, C., e Wang, L. (2019). Bigpatent: A large-scale dataset for abstractive and coherent summarization. *arXiv preprint arXiv:1906.03741*.

Simonyan, K. e Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Slamet, C., Atmadja, A., Maylawati, D., Lestari, R., Darmalaksana, W., e Ramdhani, M. A. (2018). Automated text summarization for indonesian article using vector space model. In *IOP Conference Series: Materials Science and Engineering*, volume 288, page 012037. IOP Publishing.

Souza, F., Nogueira, R., e Lotufo, R. (2020). Bertimbau: pretrained bert models for brazilian portuguese. In *Intelligent Systems: 9th Brazilian Conference, BRACIS 2020, Rio Grande, Brazil, October 20–23, 2020, Proceedings, Part I 9*, pages 403–417. Springer.

Srikanth, A., Umasankar, A. S., Thanu, S., e Nirmala, S. J. (2020). Extractive text summarization using dynamic clustering and co-reference on bert. In *2020 5th International Conference on Computing, Communication and Security (ICCCS)*, pages 1–5.

Sun, Y., Wang, S., Li, Y., Feng, S., Tian, H., Wu, H., e Wang, H. (2020). Ernie 2.0: A continual pre-training framework for language understanding. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 8968–8975.

Sutskever, I., Vinyals, O., e Le, Q. V. (2014). Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27.

Thoppilan, R., De Freitas, D., Hall, J., Shazeer, N., Kulshreshtha, A., Cheng, H.-T., Jin, A., Bos, T., Baker, L., Du, Y., et al. (2022). Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., e Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

Vinyals, O., Toshev, A., Bengio, S., e Erhan, D. (2015). Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164.

Wagner Filho, J. A., Wilkens, R., Idiart, M., e Villavicencio, A. (2018). The brwac corpus: a new open resource for brazilian portuguese. In *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*.

Wang, J., Ma, S., e Zhang, C. (2017). Citationas: A summary generation tool based on clustering of retrieved citation content. *Framework*, 7(8):19–27.

Wang, Q., Downey, C., Wan, L., Mansfield, P. A., e Moreno, I. L. (2018). Speaker diarization with lstm. In *2018 IEEE International conference on acoustics, speech and signal processing (ICASSP)*, pages 5239–5243. IEEE.

Xiao, W. e Carenini, G. (2019). Extractive summarization of long documents by combining global and local context. *arXiv preprint arXiv:1909.08089*.

Yang, S., Lu, W., Zhang, Z., Wei, B., e An, W. (2016). Amplifying scientific paper's abstract by leveraging data-weighted reconstruction. *Information Processing & Management*, 52(4):698–719.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R., e Le, Q. V. (2019). Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.

Yasunaga, M., Kasai, J., Zhang, R., Fabbri, A. R., Li, I., Friedman, D., e Radev, D. R. (2019). Scisummnet: A large annotated corpus and content-impact models for scientific paper summarization with citation networks. *CoRR*, abs/1909.01716.

Zolotareva, E., Tashu, T. M., e Horváth, T. (2020). Abstractive text summarization using transfer learning. In *ITAT*, pages 75–80.

A
Material do Estudo

O material do estudo consistiu num Termo de Consentimento Livre e Esclarecido (TCLE – seção A.1), um questionário de caracterização do perfil do participante (seção A.2) e um questionário para captura de opiniões sobre os resumos (seção A.3).

A.1
Termo de Consentimento Livre e Esclarecido

Natureza da Pesquisa

Eu, Dayson Nascimento, pesquisador responsável pela pesquisa "Avaliação de resumos gerados por humanos e por inteligência artificial", sob coordenação do Professor Hélio Côrtes Viera Lopes, do Departamento de Informática da PUC-Rio, lhe convido a participar como voluntário(a) neste estudo.

Nossa pesquisa visa criar um processo para resumir artigos científicos em português. Dessa forma, convidamos você a contribuir para a avaliação de *abstracts* gerados por humanos e por nossa ferramenta. Vale ressaltar que a participação nesse estudo é inteiramente opcional.

Riscos e desconfortos

Identificamos alguns riscos mínimos associados à participação nesta pesquisa e traçamos estratégias para mitigá-los:

- Desconforto físico: cansaço ou aborrecimento, caso a sessão seja longa. Vamos minimizar o tempo necessário à realização das atividades, focando nas questões mais relevantes ao objetivo do estudo.
- Quebra da segurança digital dos dados armazenados. Os dados coletados serão armazenados em ambiente seguro (mídia ou máquina sem acesso à internet ou em área protegida por senha). Além disto, o material coletado será desassociado da sua identidade, para garantir o seu anonimato e privacidade.
- Qualquer tipo de incômodo ou constrangimento. Você pode interromper a pesquisa a qualquer momento e sem qualquer prejuízo, penalização ou

constrangimento. Em nenhum lugar ficará registrado que você iniciou sua participação no estudo e optou por interrompê-la.

Garantia de anonimato, privacidade e sigilo dos dados

Esta pesquisa se pauta no respeito à privacidade, ao sigilo e ao anonimato dos participantes. Todos os dados pessoais serão acessados somente pelos pesquisadores envolvidos nesta pesquisa e anonimizados para análise ou divulgação. Já as anotações realizadas pelos participantes da pesquisa serão processadas para gerar dados agregados para análise da qualidade dos textos, sem nenhum tipo de rastreabilidade aos participantes do estudo.

Divulgação dos resultados

Os dados agregados e análises realizadas poderão ser publicados em publicações científicas e didáticas. Ao divulgarmos os resultados da pesquisa, nos comprometemos em preservar seu anonimato e privacidade, ocultando toda informação que possa revelar sua identidade. As informações brutas coletadas não serão divulgadas, apenas os dados agregados e processados.

Liberdade de recusa, interrupção, desistência e retirada de consentimento

Sua participação nesta pesquisa é voluntária. Sua recusa não trará nenhum prejuízo a você, nem à sua relação com os pesquisadores ou com a universidade. A qualquer momento você pode interromper ou desistir da pesquisa, sem que incorra nenhuma penalização ou constrangimento. Você não precisará justificar ou informar o motivo da interrupção ou desistência. Caso você mude de ideia sobre seu consentimento durante a sessão de estudo, basta comunicar sua decisão aos pesquisadores responsáveis, que então descartarão seus dados.

Consentimento

Eu, participante abaixo assinado(a), confirmo que:

1. Recebi informações detalhadas sobre a natureza e objetivos da pesquisa descrita neste documento e tive a oportunidade e esclarecer eventuais dúvidas;

2. Estou ciente de que minha participação é voluntária e posso abandonar o estudo a qualquer momento, sem fornecer qualquer razão e sem que haja quaisquer consequências negativas. Além disto, caso eu não queira responder a uma ou mais questões, tenho liberdade para isto;

3. Estou ciente de que minhas respostas serão mantidas confidenciais. Entendo que meu nome não será associado aos materiais de pesquisa e não será identificado nos materiais de divulgação que resultem da pesquisa;

4. Estou ciente de que a minha participação não acarretará qualquer ônus e que as atividades previstas na pesquisa não representam nenhum risco para mim ou para qualquer outro participante;

5. Estou ciente de que sou livre para consentir ou não com a pesquisa, conforme as opções que marco abaixo:

Concorda em participar da pesquisa?

– Concordo

– Não concordo

A.2
Questionário de Caracterização do Perfil do Participante

Qual seu nome?

Esse dado poderá ser utilizado para entrarmos em contato para ouvirmos seu feedback sobre nossa pesquisa.

Qual seu e-mail?

Esse dado poderá ser utilizado para entrarmos em contato para ouvirmos seu feedback sobre nossa pesquisa.

Qual a sua faixa etária?

- Até 17 anos
- de 18 a 24 anos
- de 25 a 35 anos
- de 36 a 50 anos
- a partir de 51 anos

Qual o seu grau de escolaridade?

- Sem escolaridade
- Fundamental Incompleto
- Fundamental Completo
- Ensino Médio Incompleto
- Ensino Médio Completo
- Ensino Superior Incompleto
- Ensino Superior Completo
- Mestrado Incompleto
- Mestrado Completo
- Doutorado Incompleto
- Doutorado Completo

A.3
Questionário para Captura de Opiniões sobre os resumos

Grupo 02

Este questionário consiste em avaliar sua percepção sobre dois resumos de texto, podendo ser escritos por humanos, uma IA ou um de cada. Leia com atenção, mantenha uma mente aberta e analise estilo, tom e padrões linguísticos. **Alguns erros com uso de letra minúscula ao invés de maiúscula podem acontecer e devem ser desconsiderados na avaliação**. Sua honestidade é essencial para o sucesso do estudo. Muito obrigado!

Resumo 01	Resumo 02
O objetivo desta pesquisa foi avaliar os conhecimentos e práticas relacionados à vigilância do desenvolvimento da criança de 160 profissionais que atuam na atenção primária à saúde, no município de Belém, Pará. Foram selecionados 40 médicos e 40 enfermeiros de unidades municipais de saúde (UMS), e 40 médicos e 40 enfermeiros do programa da família saudável (pfs). Na avaliação dos conhecimentos por meio da aplicação de teste objetivo, o percentual de acerto foi de 63,7% para médicos das umss , 57,3% para médicos do pfs , 62,1% para os enfermeiros do pfs e 54,3% para enfermeiros das umss. Na avaliação das práticas, apenas 21,8% das mães informaram que foram indagadas sobre o desenvolvimento dos seus filhos, 27,6% que o profissional perguntou ou observou o desenvolvimento da sua criança e 14,4% que receberam orientação sobre como estimulá-las. Concluímos que médicos e enfermeiros da atenção primária no município de Belém apresentam deficiências nos conhecimentos sobre desenvolvimento infantil e que a vigilância do desenvolvimento não é realizada de forma satisfatória, sendo necessárias sensibilização e capacitação dos profissionais para esta prática.	Este estudo tem como objetivo avaliar o desenvolvimento do programa de estimulação precoce no desenvolvimento infantil na atenção primária à saúde do filho, na cidade de Belém, Pará, Brasil. Trata-se de um estudo de avaliação, com dados coletados por meio de entrevista semiestruturada com 20 mães de crianças menores de seis meses. Os dados foram analisados por meio do programa estatístico de avaliação do desenvolvimento infantil e do conhecimento do filho. Constatou-se que existe diferença estatística significante entre os médicos em todas as categorias avaliadas. Nos profissionais a vigilância do desenvolvimento foi direcionada pelos médicos e enfermeiros. Verificou-se que 92% das mães não atingiram um nível de conhecimento sobre o desenvolvimento infantil. Quanto à sua prática diária, 39% dos profissionais buscam orientar os médicos para uma área de saúde, o que demonstra uma lacuna entre estes profissionais. Observou-se, ainda, a necessidade de a capacitação das mães para uma mudança nos serviços básicos no modelo de atenção à saúde da criança.

Figura A.1: Comparação entre Resumos 01 e 02

*
Avaliação resumo 01:

	Discordo totalmente	Discordo	Discordo Parcialmente	Não concordo nem discordo	Concordo Parcialmente	Concordo	Concordo totalmente
Considero que o resumo tenha uma boa estrutura lógica (possui objetivo, metodologia e resultados)	○	○	○	○	○	○	○
Considero que eu consegui entender o objetivo principal do resumo	○	○	○	○	○	○	○
Considero que o resumo tenha sido elaborado por um ser humano	○	○	○	○	○	○	○

Figura A.2: Avaliação do Primeiro Resumo

*
Avaliação resumo 02:

	Discordo totalmente	Discordo	Discordo Parcialmente	Não concordo nem discordo	Concordo Parcialmente	Concordo	Concordo totalmente
Considero que o resumo tenha uma boa estrutura lógica (possui objetivo, metodologia e resultados)	○	○	○	○	○	○	○
Considero que eu consegui entender o objetivo principal do resumo	○	○	○	○	○	○	○
Considero que o resumo tenha sido elaborado por um ser humano	○	○	○	○	○	○	○

Figura A.3: Avaliação do Segundo Resumo

*
Avaliação geral:

	Discordo totalmente	Discordo	Discordo Parcialmente	Não concordo nem discordo	Concordo Parcialmente	Concordo	Concordo totalmente
Considero que os dois resumos acima são semelhantes em conteúdo	○	○	○	○	○	○	○

Figura A.4: Similaridade entre Resumos. As respostas padronizadas contêm 7 opções na Escala de Likert.