XXIV ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO – XXIV ENANCIB

ISSN 2177-3688

GT 8 – Informação e Tecnologia

INTELIGÊNCIA ARTIFICIAL EM CIÊNCIA DA INFORMAÇÃO: UMA PROPOSTA DE FERRAMENTA EM MODELO BERT PARA ORGANIZAÇÃO E RECUPERAÇÃO DA INFORMAÇÃO DAS PUBLICAÇÕES DA ISKO BRASIL

***ARTIFICIAL INTELLIGENCE IN INFORMATION SCIENCE**: A PROPOSAL FOR A BERT MODEL TOOL FOR ORGANIZING AND RETRIEVING INFORMATION FROM ISKO BRASIL PUBLICATIONS*

Rita do Carmo Ferreira Laipelt – Universidade Federal do Rio Grande do Sul (UFRGS)
Simone Dias Marques – Universidade Federal do Rio Grande do Sul (UFRGS)

Modalidade: Resumo expandido

Resumo: Apresenta-se a aplicação do modelo de Inteligência Artificial BERT para a organização e recuperação da informação em *corpus* de documentos em PDF publicados pela International Society for Knowledge Organization (ISKO) Brasil. Objetivo: Propõe-se a adaptação do algoritmo BERTimbau, cujo código é aberto, para melhorar a acessibilidade e recuperação de informações nas publicações da ISKO Brasil a partir da extração automática de termos-chave. A abordagem metodológica integra a perspectiva da Análise de Domínio e pesquisa aplicada. Resultados: A aplicação extraiu 100 termos, apontando para a necessidade refinamento de etapas para o treino deste modelo de IA em desenvolvimento.

Palavras-chave: Recuperação da Informação; Organização do Conhecimento; inteligência artificial; processamento de linguagem natural (NLP).

Abstract: This article presents the application of the BERT Artificial Intelligence model for the organization and retrieval of information in a corpus of PDF documents published by the International Society for Knowledge Organization (ISKO) Brazil. Objective: It proposes the adaptation of the BERTimbau algorithm, which is open-source, to improve accessibility and information retrieval in ISKO Brazil's publications through the automatic extraction of key terms. The methodological approach integrates the perspective of Domain Analysis and applied research. Results: The application extracted 100 terms, indicating the need for refinement steps in training this developing AI model.

Keywords: Information Retrieval; Knowledge Organization; artificial intelligence; natural language processing (NLP).

1 INTRODUÇÃO

Atualmente, diante da exponencialidade informacional na Web, ao fazer uma busca em repositórios científicos, é comum encontrar resultados repetidos. Este formato torna a pesquisa morosa e dificulta a apropriação social, científica e tecnológica da informação e, em última instância, ao construto do conhecimento. Além disso, por se basear em indexação de palavras-chave, sistemas de recuperação de informação na Web não costumam permitir a exploração contextualizada dos termos, informando simplesmente sobre a existência ou não e a localização de documentos relacionados à sua requisição. Essa lacuna provoca a reflexão sobre o desenvolvimento de soluções para aprimorar a recuperação da informação, sobretudo diante dos avanços em Inteligência Artificial (IA).

No entanto, o maior obstáculo para a pesquisa a respeito do funcionamento dos modelos generativos de IA é saber como essas ferramentas operam e são treinadas, pois seus códigos e conjuntos de dados são obscuros. Este aspecto "caixa-preta" não é somente de nível técnico, mas abrange aspectos políticos, éticos e culturais. É neste horizonte que este estudo se insere, com o objetivo de contribuir para possibilidades de aplicação de IA na Recuperação da Informação utilizando o modelo BERT (Bidirectional Encoder Representations from Transformers), uma arquitetura de linguagem natural em código aberto que tem sido aprimorada desde seu lançamento em 2018.

Nossa pesquisa[*] propõe a aplicação de modelo BERT no *corpus* de publicações online da International Society for Knowledge Organization (ISKO) Brasil. Para tanto, criou-se uma ferramenta piloto desenvolvida a partir do código disponibilizado pela Open Knowledge Brasil, a qual utiliza um algoritmo chamado "BERTimbau" para classificar, contextualizar e trazer visibilidade às informações presentes em documentos em formato PDF. Assim, o caráter descritivo e experimental do estudo sugere a possibilidade de adaptação deste modelo como uma ferramenta de IA para a recuperação e organização da informação. No primeiro teste aqui descrito, optamos por aplicar a ferramenta a somente um dos documentos, escolhido pelo

* Agradecemos ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) pelo apoio financeiro ao Projeto de Pesquisa 421849/2023-1, o qual ajudou a viabilizar nossa participação no XXIV Enancib.

critério de menor extensão de páginas.

Smiraglia (2011) enfatiza que a organização do conhecimento depende da identificação e ordenação dos conceitos a partir do mapeamento de conteúdos ou heurísticas em domínios específicos. Por sua vez, Hjørland e Albrechtsen (1995) afirmam que a análise de domínio é uma abordagem que visa compreender a estrutura e a dinâmica do conhecimento dentro de um campo específico, permitindo identificar as inter-relações entre conceitos e as necessidades informacionais dos usuários.

Nesse sentido, entendemos que a Análise de Domínio é a perspectiva teórico-metodológica de primeira escolha para identificar os principais conceitos e termos específicos no *corpus* em análise, auxiliando na compreensão dos requisitos conceituais e do entendimento do contexto utilizado na comunidade de discurso da ISKO Brasil e do domínio da Ciência da Informação. Igualmente, a AD contribui para o mapeamento de elementos-chave para compor a terminologia específica e a classificação de *tokens* no treino do algoritmo. Conforme a evolução dos testes, esperamos contribuir para a acessibilidade, legibilidade e recuperação de informações nas publicações da ISKO Brasil e facilitar a construção de outros sistemas de organização do conhecimento a partir de seu potencial de replicação a diversos contextos da Ciência da Informação.

2 DESENVOLVIMENTO

Fundada em 1989, a International Society for Knowledge Organization (ISKO) é a principal sociedade científica responsável pela área de Organização do Conhecimento. Com um escopo amplo e interdisciplinar, tem a missão de incentivar o desenvolvimento de trabalhos conceituais sobre a organização do conhecimento em todas as suas formas, como por exemplo, banco de dados, bibliotecas, dicionários e Internet. Reúne profissionais de diferentes áreas nos campos da Ciência da Informação, Filosofia, Linguística e Ciência da Computação.

A série de publicações “Estudos Avançados em Organização e Representação do Conhecimento” da ISKO Brasil contém os anais completos dos congressos nacionais realizados pela entidade e visa divulgar as pesquisas e práticas em organização do conhecimento no Brasil

e no mundo. A série é composta por 12 volumes, publicados entre 2009 e 2020, cada um com ISBN e DOI. Tais publicações tratam de temas valiosos à Ciência da Informação e à pesquisa científica, como privacidade, pós-verdade, datificação, ontologias, taxonomias, linguagens documentárias, entre outros. Em formato digital (PDF), são 7 volumes disponíveis com acesso gratuito no site da ISKO Brasil.

Apesar da possibilidade de *download*, o conteúdo destes documentos está inacessível para os usuários que realizam pesquisa na Web. Igualmente, não é possível fazer uma busca por termos no próprio site, pois não há um mecanismo interno de busca, o que inviabiliza a localização de artigos sem abrir cada volume manualmente. Ademais, o formato PDF também apresenta obstáculos à organização e recuperação da informação, pois não permite a extração de dados facilmente, o que impede a indexação da informação com outros sistemas, repositórios ou plataformas, dificultando a apropriação social da informação. Portanto, uma ferramenta de IA capaz de classificar, indexar e encontrar documentos a partir da terminologia específica da OC pode resolver estes problemas de acessibilidade, legibilidade e de busca e recuperação das informações das publicações da ISKO Brasil.

Para tanto, construímos um piloto da ferramenta no formato de um mínimo produto viável *Minimum Viable Product - MVP*). Assim, foi possível validar uma versão simples antes de investir recursos no seu desenvolvimento. Aplicamos neste piloto as funcionalidades básicas necessárias para testar a abordagem de extração de termos usando o modelo de BERTimbau em um documento escolhido[1]. A partir da fase de teste, acredita-se que é possível fazer ajustes e refinamentos no modelo, ajustar seus parâmetros e melhorar a precisão da ferramenta.

2.2 Sobre os modelos BERT e BERTimbau

Um dos avanços mais notáveis em Inteligência Artificial no Processamento de Linguagem Natural (NLP) é o modelo BERT (Bidirectional Encoder Representations from

[1] O artigo escolhido está contido na publicação "Organização do Conhecimento em diferentes contextos: desafios e perspectivas na era da datificação". Disponível em https://isko.org.br/wp-content/uploads/2023/06/livro-isko-Brasil_23.pdf, pp. 261-270.

Transformers). Introduzido por pesquisadores do Google AI Language em 2018, o BERT revolucionou a forma como as máquinas entendem a linguagem humana, levando a melhorias na precisão e relevância dos resultados dos mecanismos de pesquisa, tradução de idiomas e outras tarefas baseadas em linguagem. Sua construção é projetada para compreender as nuances e o contexto da linguagem humana para que os usuários possam interagir de forma mais natural com a tecnologia. É o primeiro sistema não supervisionado e bidirecional para pré-treinamento de modelos de processamento de linguagem natural.

Através de sua bidirecionalidade, o BERT examina o contexto de uma palavra observando aquelas que vêm antes e depois dela – algo que não era possível com modelos anteriores e até mesmo hoje, com os modelos massivos de conversa GPT, (Generative Pre-trained Transformers) que normalmente utilizam estatística de probabilidades para "adivinhar" uma palavra na direção posterior e gerar aquela que é mais frequente em determinado enunciado e idioma.

Por sua vez, e de forma resumida, o modelo BERT é baseado na arquitetura Transformer, que usa um mecanismo que aprende relações contextuais entre palavras (ou subpalavras) em um texto. Além disso, é capaz de entender as relações semânticas de um corpus textual e representá-lo. Ao contrário dos modelos anteriores que analisavam sentenças da esquerda para a direita ou da direita para a esquerda, o BERT multilíngue é pré-treinado em um grande *corpus* de texto não rotulado, incluindo toda a Wikipedia (2.500 milhões de palavras) e o Book Corpus (800 milhões de palavras).

Nas palavras de Wang *et al.* (2024, p. 7), os "modelos baseados em BERT são especificamente adequados em tarefas de recuperação de informações para aplicações como mecanismos de pesquisa" devido à sua compreensão semântica eficaz e capacidade de contextualização até mesmo em documentos longos. Entretanto, a necessidade de uma capacidade computacional robusta e enormes quantidades de dados não rotulados, com relatos de modelos sendo treinados usando milhares de GPUs ou TPUs e centenas de gigabytes de dados textuais brutos, limitavam as aplicações do modelo e a sua disponibilidade em idiomas como o português (Souza *et. al*, 2020).[2]

[2] Souza *et al*. (2020) explicam que muito esforço foi dedicado ao pré-treinamento de BERT monolíngue e modelos derivados de BERT em idiomas únicos, como francês, holandês, espanhol,

Surgiram assim modelos "pequenos", em relação aos Large Language Models (LLMs), como o BERTimbau, o AlBERT, o Albertina e outros. Para os fins da pesquisa que relatamos neste artigo, focaremos no BERTimbau. Desenvolvido pela Neuralmind AI, BERTimbau é um modelo BERT pré-treinado em um grande *corpus* de textos em português brasileiro que alcança desempenhos julgados eficazes de reconhecimento de entidade nomeada, similaridade textual de frase e reconhecimento de implicação textual em língua portuguesa do Brasil (Souza *et al*. 2020). A escolha do modelo BERT em vez de modelos GPT para esta pesquisa é fundamentada em considerações técnicas e de desempenho, especialmente em relação à análise semântica, como sumarizado na Tabela 1 abaixo.

Tabela 1 – Comparação entre BERT e GPT para análise semântica para o idioma português

Critério	Descrição
Análise semântica acurada	O BERT entende o contexto bidirecional das palavras, resultando em uma análise semântica mais precisa. Quanto mais específico o *corpus*, melhor é sua eficácia (Souza *et. al*, 2020).
Arquitetura bidirecional	O BERT processa o texto de forma bidirecional, capturando relações complexas entre palavras e contexto, ao contrário dos modelos GPT, que são unidirecionais e são utilizados para **gerar** respostas.
Aplicações específicas	O BERTimbau, versão do BERT treinada para o português, é mais eficaz em lidar com as nuances da língua do que os modelos GPT, que são geralmente treinados em múltiplos idiomas.

Fonte: Elaborada pelos autores, 2024.

Uma vez que a base de código de BERTimbau é aberta e está disponível para ser modificada e adaptada, escolhemos este modelo de IA para a aplicação em um MVP voltado para facilitar a busca de documentos da ISKO Brasil. Sua versatilidade permite não apenas personalizar a ferramenta para extrair termos específicos do campo da organização do conhecimento como permitir a evolução contínua do modelo conforme surgirem novas necessidades, sem custos adicionais significativos. Além disso, pode ser utilizado também em

italiano e outros devido ao desempenho superior e eficiência de recursos computacionais em comparação ao BERT multilíngue.

outras pesquisas que demandem a identificação de terminologias específicas, na extração de termos e na análise de documentos em diferentes domínios.

2.1 Metodologias

I Análise de Domínio: Para examinar e mapear as relações conceituais presentes nos documentos da ISKO Brasil, este estudo utiliza a análise de domínio, que permite identificar os conceitos centrais e suas interconexões, facilitando a representação ontológica dos termos relevantes. Conforme destacado por Smiraglia (2011), a análise de domínio se caracteriza pelo exame dos aspectos teóricos de um determinado contexto, representado por uma literatura específica ou uma comunidade de pesquisadores. Nesse contexto, a Teoria do Conceito de Dahlberg (1978) oferece uma estrutura metodológica adicional para a modelagem de domínios, ao estabelecer categorias que agrupam e estruturam os conceitos representativos de um dado domínio, possibilitando a compreensão das relações semânticas e pragmáticas entre os conceitos para a organização e a recuperação do conhecimento em sistemas de informação.

II Pesquisa Aplicada[3]: Desenvolvimento e teste do protótipo de IA utilizando Python e o modelo BERTimbau para reconhecer e extrair conteúdos de documentos PDF conforme a terminologia utilizada pela comunidade de colaboradores da ISKO Brasil, contribuindo para a recuperação da informação de forma precisa e contextualizada.

Neste primeiro teste da ferramenta, criamos um dicionário vazio chamado *terminology* para armazenar a terminologia. Lemos o documento escolhido em PDF e identificamos os termos relevantes usando uma função <extract_terms()>. Para cada termo, adicionamos ao nome do arquivo PDF a uma lista associada a esse termo no dicionário *terminology*. Definimos também uma função *search*() que recebe uma consulta de pesquisa. Se o termo estiver no dicionário *terminology*, a lista de arquivos PDF associados a esse termo retorna. Caso contrário, retorna uma lista vazia. O teste foi realizado no Google Colab através do navegador utilizando a biblioteca aberta Hugging Face Transformers.

[3] Por conta do limite de número de páginas deste artigo, deixaremos de inserir o código-fonte cujo modelo genérico está disponível na plataforma Github, como consta nas referências.

3 RESULTADOS

A partir desta abordagem em BERTimbau, que adaptou o código já existente no repositório aberto Github, obtivemos uma lista de 100 termos extraídos do documento escolhido. Deste total, 50 foram considerados repetidos/similares ou irrelevantes (verbos, adjetivos), pois este protótipo ainda não passou por *fine tuning* e outras etapas de refinamento. Observamos ainda que ao rodar, o *bot* "escolheu" certos termos do Índice, ou seja, ele não os extraiu dos artigos. Por fim, a função *search*() retornou erro.

Estas limitações apontaram para a necessidade de redeterminar as funções necessárias no código e no modelo, tais como especificaremos a seguir. Eis alguns dos 100 termos extraídos após implementadas as regras no sistema:

Quadro 1 – Termos extraídos do documento com o código adaptado do BERTimbau.

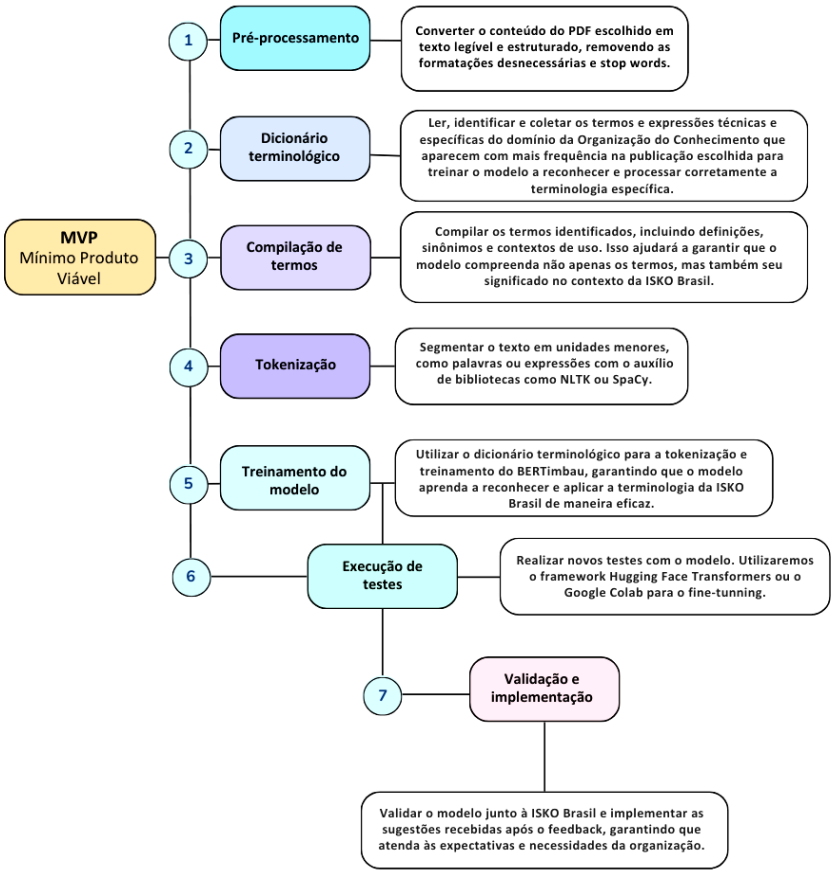
Termo	Termo	Termo
Organização do Conhecimento	Repositório de Dados	Datificação
Representação do Conhecimento	Dados Abertos	Memória
Tecnologia	Processamento de Linguagem	Cultura
Diversidade Cultural	Gestão Documental	Ontologias
Sistema de Conhecimento	Chatbots	Classificação
Contextualização	Inteligência Artificial	Objetos de Fronteira
Tradução terminológica	Capacitismo	Indexação Arquivística
Big data	Estratégias de Preservação	Intencionalidade
Expansão da Informação	Necessidades de Informação	Domínios de Especialidade
Ética da Informação		

Fonte: Elaborado pelos autores (2024).

A partir dos resultados deste primeiro teste, evidenciou-se a necessidade de realização

de novas etapas de refinamento do modelo, a fim de processar e extrair informações relevantes desses documentos (por exemplo, termos, conceitos, temas, palavras-chave) e treinar o algoritmo conforme o fluxograma da Figura 1 a seguir:

Figura 1 – Fluxograma de etapas para treino do BERTimbau em documento PDF.



Fonte: Elaborado pelos autores (2024).

A expectativa é de que estas etapas garantam que o MVP em BERTimbau atenda aos objetivos de extrair e processar informações relevantes da publicação escolhida da ISKO Brasil, permitindo que o algoritmo aprenda de maneira eficaz a partir da terminologia específica do domínio. Além disso, o sistema poderá ser escalável conforme necessidades do acervo de publicações.

4 CONSIDERAÇÕES FINAIS

A Recuperação da Informação (RI) é um dos eixos epistemológicos da Ciência da Informação, envolvendo aspectos teóricos e práticos que facilitam a busca e a apropriação social do conhecimento em um corpus discursivo (Barros; Laipelt, 2021). Saracevic (1992) define RI como os aspectos intelectuais da descrição da informação e os sistemas utilizados para sua busca, relacionando-se com diversos campos do conhecimento.

A aproximação entre a Ciência da Informação e a Ciência da Computação se intensificou a partir dos anos 1990, impulsionada pela crescente digitalização da informação e pelo desenvolvimento da Web, redes e sistemas de informação. Hoje, com o avanço das inteligências artificiais treinadas para compreensão de linguagem natural, acreditamos que essas ferramentas possam ser adotadas e aprimoradas como instrumentos de organização do conhecimento na Ciência da Informação.

Propomos, assim, que a implementação da ferramenta BERTimbau no âmbito das publicações da ISKO Brasil possa contribuir para aumentar a visibilidade de documentos valiosos à Ciência da Informação. Por ser um artefato de código aberto, pode ser também adaptado para otimizar o desenvolvimento de Sistemas de Organização do Conhecimento (SOCs), facilitar a construção de tesauros, ontologias e outros recursos que exigem uma representação precisa de conceitos e suas inter-relações. Consideramos que esta pesquisa possibilitará a evolução do modelo aqui proposto para ser replicável e abrir caminhos futuros para escalar a ferramenta a outros contextos.

REFERÊNCIAS

BARROS, Thiago H. B.; LAIPELT, Rita do Carmo F. Uma análise de domínio da área de Organização e Representação do Conhecimento no contexto do periódico *Em Questão*. **Em Questão**, Porto Alegre, v. 27, n. 4, p. 438-468, out./dez. 2021. DOI: http://dx.doi.org/10.19132/1808-5245274.438-468.

DAHLBERG, Ingetraut. Teoria do conceito. Tradução: Astério Tavares Campos. **Ciência da Informação**, Rio de Janeiro, v. 7, n. 2, p. 101-107, 1978c.

HJØRLAND, Birger. The domain-analytic approach to Knowledge Organization and the case of Library and Information Science. **Journal of Documentation**, v. 60, n. 6, p. 582–605, 2004.

HJØRLAND, Birger. Domain analysis: a sociocognitive orientation for Information Science research. **Bulletin of the American Society for Information Science and Technology**, v. 30, n. 3, p. 17-21, 2004.

HJØRLAND, Birger. Fundamentals of Knowledge Organization. **Knowledge Organization**, v. 30, n. 2, p. 87-111, 2003.

HJØRLAND, Birger; ALBRECHTSEN, H. Toward a new horizon in information science: domain-analysis. **Journal of the American Society for Information Science**, v. 46, n. 6, p. 400-425, 1995.

OPEN KNOWLEDGE BRASIL. **Querido Diário**: 1º Relatório Técnico de Atividades. Documentação, 2021. Disponível em: https://www.ok.org.br/wp-content/uploads/2021/07/Querido-Diario-1o-Relatorio-Tecnico-de-Atividades.pdf.

SARACEVIC, T. Information science: origin, evolution and relations. In: VAKKARI, Pertti; CRONIN, Blaise (Ed.). **Conceptions of library and information science**: Historical, empirical, and theoretical perspectives. London: Taylor Graham, 1992. p. 2.

SMIRAGLIA, R. P. **The elements of knowledge organization.** Cham: Springer, 2014.

SOUZA, Fábio; NOGUEIRA, Rodrigo; LOTUFO, Roberto. BERT models for Brazilian Portuguese: Pretraining, evaluation and tokenization analysis. **Applied Soft Computing**, v. 149, Part A, 2023. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S1568494623009195. Acesso em: 18 mar. 2024.

SOUZA, Fábio; NOGUEIRA, Rodrigo; LOTUFO, Roberto. BERTimbau: Pretrained BERT Models for Brazilian Portuguese. *In*: INTELLIGENT SYSTEMS: Brazilian Conference, 9., Rio Grande, 2020. **Proceedings**: Part I. Berlin: Springer-Verlag, 2020. p. 403–417. ISBN 978-3-030-61376-1.

WANG, *et al*. **Utilizing BERT for information retrieval**: survey, applications, resources, and challenges. Disponível em: https://ar5iv.labs.arxiv.org/html/2403.00784v1. Acesso em: 14 jul. 2024.