AKR *Advances on Knowledge Representation Journal*

O uso do ChatGPT na indexação dos pedidos de Acesso à Informação no Fala.BR

Zenóbio dos Santos Júnior[1], Frederico Giffoni de Carvalho Dutra[2], Eduardo José da Silva Luz[3]

[1]Doutor em Tecnologia da Informação e Comunicação e Gestão do Conhecimento – Universidade FUMEC

[2]Professor do Programa de Pós-Graduação em Tecnologia da Informação e Comunicação e Gestão do Conhecimento – Universidade FUMEC

[3]Professor Adjunto no Departamento de Computação – Universidade Federal de Ouro Preto

zenojr19@gmail.com, fgcdutra@gmail.com, eduluz@ufop.edu.br

Resumo: *Este estudo teve por objetivo analisar a indexação dos pedidos de acesso às informações das Universidades Federais do Brasil com intuito de propor melhorias, por meio do ChatGPT, no Fala.BR. A pesquisa estruturou-se a partir dos conceitos de acesso à informação, classificação e inteligência artificial para a indexação dos Assuntos, Subassuntos e Tags no Fala.BR. Trata-se de uma pesquisa exploratória e explicativa, de natureza aplicada, cuja abordagem é qualitativa-quantitativa e os procedimentos técnicos utilizados foram a bibliográfica e experimental tendo como amostra os usuários do Fala.BR e Arquivistas. Utilizaram-se instrumentos de pesquisa como o vocabulário controlado do governo eletrônico e o código de classificação de documentos de arquivo das atividades fim das Instituições Federais de Ensino Superior para geração de prompts no ChatGPT. Em seguida, os resultados foram comparados com os textos dos arquivistas que utilizaram apenas o*

[1] Doutor e Mestre em Tecnologia da Informação e Comunicação e Gestão do Conhecimento pela Universidade FUMEC; Bacharel em Arquivologia pela Universidade Federal Fluminense; Arquivista e Coordenador do Arquivo Central da Universidade Federal de Ouro Preto. ORCID: https://orcid.org/0000-0003-1380-7391

[2] Professor, Pesquisador, Doutor e Mestre em Gestão da Informação e do Conhecimento pela Universidade Federal de Minas Gerais (2020), Especialista em Gestão Estratégica de Marketing (2007) e Graduado em Administração (2005). Atua na área de Comunicação, Marketing e Inteligência da Companhia Energética de Minas Gerais - CEMIG, com foco em inteligência e monitoramento de clientes e marcas nas redes sociais e leciona em cursos de graduação e pós-graduação. Universidade FUMEC. ORCID: https://orcid.org/0000-0002-8666-0354

[3] Doutor em Ciência da Computação pela Universidade Federal de Ouro Preto (UFOP) em 2019, onde seu projeto de pesquisa foi reconhecido com o IBM Ph.D. Fellowship Award (2017). Professor Adjunto no Departamento de Computação (DECOM) da Universidade Federal de Ouro Preto. Possui experiência em pesquisa científica e desenvolvimento tecnológico, com foco em Inteligência Artificial, Redes Neurais Artificiais, Visão Computacional e Reconhecimento de Padrões. ORCID: https://orcid.org/0000-0001-5249-1559

AKR Advances on Knowledge Representation Journal

código de classificação e, mesmo assim, chegaram a números surpreendentes nos campos Subassuntos e Tags. Ao ponto de inverter os valores desses campos que estavam negativos, devido à ausência de preenchimento nos pedidos de acesso à informação originais, passou a ser positivo devido ao grau de similaridade de concordância substancial entre o homem e a máquina.

The use of ChatGPT in indexing Access to Information requests on Fala.BR

***Abstract**: This study aimed to analyze the indexing of requests for access to information from Federal Universities in Brazil with the aim of proposing improvements, through ChatGPT, on Fala.BR. The research was structured based on the concepts of access to information, classification and artificial intelligence for indexing Subjects, Sub-subjects and Tags in Fala.BR. This is an exploratory and explanatory research, of an applied nature, whose approach is qualitative-quantitative and the technical procedures used were bibliographic and experimental, using Fala.BR and Arquivistas users as a sample. Research instruments such as the controlled e-government vocabulary and the classification code for archival documents of the core activities of Federal Higher Education Institutions were used to generate prompts in ChatGPT. Then, the results were compared with the texts of archivists who used only the classification code and, even so, reached surprising numbers in the Subsubjects and Tags fields. To the point of inverting the values of these fields, which were negative, due to the absence of completion in the original information access requests, they became positive due to the degree of similarity of substantial agreement between man and machine.*

1. Introdução

Atualmente, muitas informações são registradas eletronicamente por vários tipos de sistemas, incluindo os de órgãos públicos. De acordo com o Programa de Governo Eletrônico Brasileiro (BRASIL, 2023a), desde sua criação, em 2000, transformou as relações do Governo com os cidadãos, empresas e os órgãos do próprio governo, de forma a aprimorar a qualidade dos serviços prestados, promover a interação com empresas e indústrias, e fortalecer a participação cidadã por meio do acesso à informação a uma administração mais eficiente.

A Lei nº 12.527/2011 garante ao cidadão o direito de acesso à informação (LAI) de duas formas: transparência ativa e passiva. Na primeira, os órgãos públicos disponibilizam um rol de informações em seus sítios (Institucional, ações e programas,

AKR Advances on Knowledge Representation Journal

licitações e contratos, servidores, informações classificadas, etc.), enquanto a segunda oferece um canal de atendimento para o cidadão denominado Fala.BR - plataforma integrada de Ouvidoria e Acesso à Informação (BRASIL, 2023b). A plataforma nasceu a partir da unificação dos sistemas de Serviço de Informações ao Cidadão (e-SIC) e de Ouvidorias (e-Ouv) em 31 de agosto de 2020 e contempla os tipos de manifestação de acesso à informação e ouvidoria (denúncia, elogio, reclamação, simplifique, solicitação e sugestão).

É importante citar que os dados desta pesquisa são resultantes dos pedidos de acesso à informação que foram extraídos do Fala.BR em 2023, a partir do Painel LAI - criado para facilitar o monitoramento e o cumprimento da LAI pelos órgãos e entidades do Poder Executivo Federal (BRASIL, 2023c); da Busca de Pedidos e Respostas - que permite a consulta pública dos pedidos de informação, direcionados aos órgãos e às entidades com as respectivas respostas fornecidas (BRASIL, 2023e); e do Download de Dados - uma base de dados dos pedidos e respostas realizados no Poder Executivo Federal, em formatos CSV e XML (BRASIL, 2023d), permitindo assim, obter os Assuntos e seus desdobramentos como os Subassuntos e as Tags. No âmbito desta pesquisa e, para melhor entendimento, o termo “Principais Temas” integrará todos os Assuntos, Subassuntos e Tags.

A identificação de lacunas dos principais temas registrados nos pedidos de acesso à informação, seja pela ausência de informações, confusão ou mesmo desconhecimento do usuário quanto ao preenchimento dos campos, possibilita uma proposição de aperfeiçoamento do sistema Fala.BR por meio dessa pesquisa.

Estudos anteriores constataram que os Principais Temas necessitavam de ajustes e melhorias no Fala.BR, como demonstrado por Santos Júnior et al (2022) nos pedidos realizados entre janeiro a abril de 2022, onde dos 178 assuntos existentes na plataforma, metade deles (49,20%) foram cadastrados como “Acesso à Informação”, sendo esse um rótulo genérico. Enquanto isso, Santos Júnior; Corrêa; De Faria (2023, p. 376) apontaram a ausência de preenchimento dos campos Subassuntos (83,47%) e Tags (76,81%) pelos usuários que utilizaram o Fala.BR durante todo o ano de 2021. Para solucionar esses problemas, buscou-se na Arquivologia e na Ciência da Computação uma estratégia para automatizar a classificação de documentos das atividades-fim das Instituições Federais de Ensino Superior (IFES). Nossa hipótese é que grandes modelos de linguagem, como o GPT-4, são capazes de classificar automaticamente um documento a partir de um vocabulário controlado, de forma a prover ou ajustar campos dos "Principais Temas" (Assunto, Subassunto e Tag). Para isso, foi construído um conjunto de dados rotulados por especialistas, a partir de pedidos de acesso à informação feitos na plataforma Fala.BR, permitindo assim a avaliação dos resultados gerados tanto pelo modelo computacional quanto por humanos.

AKR Advances on Knowledge Representation Journal

Diante do exposto, surge a seguinte questão: Quais as contribuições o ChatGPT pode trazer para a indexação dos pedidos de acesso à informação das Universidades Federais, no Fala.BR?

Dessa forma, o objetivo deste artigo é propor o uso do ChatGPT na indexação dos pedidos de Acesso à Informação das Universidades Federais do Brasil no Fala.BR, a partir da utilização do plano de classificação de documentos da atividade-fim das Instituições Federais de Ensino Superior (IFES).

Esta pesquisa está estruturada em 5 (cinco) partes. Além desta introdução, a seção seguinte destaca os pilares teóricos que sustentam esta investigação. Por conseguinte, os procedimentos metodológicos, que regem esta pesquisa, são descritos e, em sequência, aplicados na análise dos resultados. Adiante, as considerações finais são articuladas e as referências que embasam esta investigação são listadas, finalizado esta pesquisa.

2. Desenvolvimento

Nesta seção, para melhor compreensão do texto, serão apresentados três tópicos relacionados ao Acesso à Informação, Classificação e ChatGPT.

2.1 Acesso à Informação

A Lei de Acesso à Informação - LAI, regulamenta o Acesso à Informação previsto na Constituição Federal de 1988 (BRASIL, 2011) e traz capítulos importantes como: divulgação das informações pelos órgãos em seus sítios na *Internet* (Transparência Ativa); e a disponibilização de ferramenta para realização de pedidos de acesso à informação e de Ouvidoria, o Fala BR (Transparência Passiva). Neste, é possível realizar e acompanhar o pedido, entrar com recursos, apresentar reclamações, denúncias etc (BRASIL, 2023b).

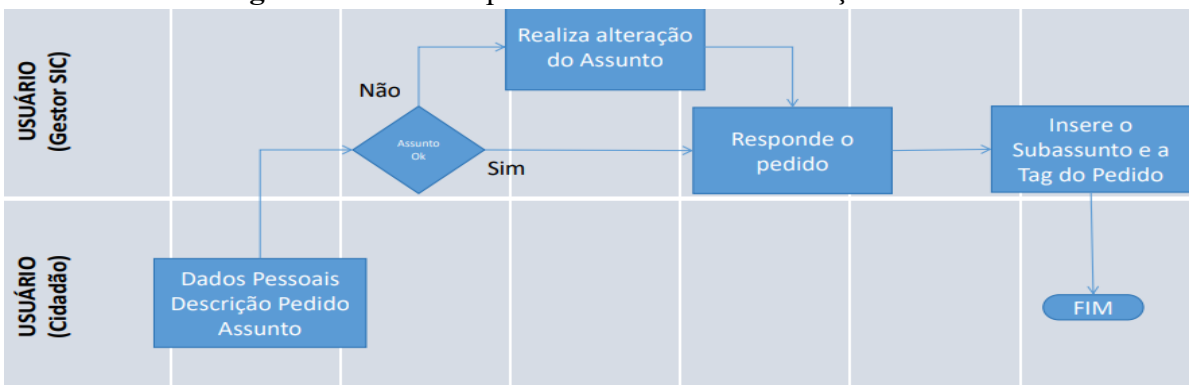
O sistema Fala.BR apresenta 21 (vinte e um) campos de cadastramento, dentre eles: Protocolo, Órgão Destinatário, Prazo de Atendimento, Forma de Resposta, Assunto, Subassunto, *Tag*, Decisão, entre outros. Ao realizar o pedido de acesso à informação, o usuário (cidadão) preenche o cadastro (dados pessoais)[4], descreve o pedido que se quer tratar e o tipo de assunto referente ao tema (campo obrigatório). O gestor do órgão, responsável pelo Serviço de Informação ao Cidadão (SIC), encaminha o pedido ao setor responsável para formulação da resposta e aguarda o retorno para finalização do atendimento. Ao contrário do Assunto, o Subassunto e a *Tag* são campos de

[4] Os pedidos de Acesso à informação podem ser realizados de forma anônima no Fala.BR (Brasil, 2023b).

AKR Advances on Knowledge Representation Journal

preenchimento opcional. A figura 1, retrata de forma resumida, o fluxo do pedido de acesso à informação dando ênfase aos principais temas.

Figura 1 - Fluxo do pedido de acesso à informação



Fonte: Elaborado pelos autores

Como mencionado, há grande incidência de um único Assunto e ausência de preenchimento dos campos Subassuntos e *Tags*. Daí a proposta é facilitar os usuários na indexação dos principais temas, contribuindo para uma pesquisa mais eficaz e recuperação da informação. Para isso, a pesquisa terá o auxílio de instrumentos técnicos de arquivologia e de inteligência artificial para agilizar o processo e beneficiar o usuário.

2.2 Classificação

Para a Arquivologia, a Classificação é definida como uma organização dos documentos de um arquivo ou coleção, de acordo com um plano de classificação, código de classificação ou quadro de arranjo (BRASIL, 2005). Também pode ser definida como um processo, sistema ou disciplina (Lima, 2021) que descreve os fundamentos teóricos e conceitos relacionados à gestão documental e a sua importância para o funcionamento de um arquivo (Do Nascimento Silva et al., 2020, p. 41); a classificação tem importância "na padronização dos termos, palavras-chave ou descritores resultantes do processo de indexação, cumprindo a função de criar rótulos que, junto com o número de classificação dos documentos, representam pontos de acesso para a recuperação dos documentos demandados pelos usuários" (Sousa; Araújo Júnior, 2017, p. 53). De acordo com Souza (2000, p. 4) "se faz presente nas duas principais funções de um sistema tradicional de recuperação de informação: na habilidade para identificar itens de informação específicos e definidos de forma precisa (sistemas de indexação de assuntos) e na habilidade de demonstrar as áreas de assunto disponíveis (estruturas classificatórias /instrumentos de classificação)".

AKR Advances on Knowledge Representation Journal

O código de classificação e tabela de temporalidade e destinação de documentos relativos às atividades-meio do Poder Executivo Federal, deixam claro que adotou-se a mecânica do sistema de classificação decimal desenvolvido por Dewey em 1873, isto é, a divisão dos conjuntos documentais de determinado órgão ou entidade, produzidos em razão de suas funções e atividades, em dez classes e, sucessivamente, em dez subclasses, dez grupos e dez subgrupos, usando-se para isso a notação decimal (BRASIL, 2020, p. 7).

No âmbito da pesquisa, as Universidades Federais, a utilização do Código de Classificação de Documentos de Arquivo (CCDA) para atividades-fim das Instituições Federais de Ensino Superior - IFES (BRASIL, 2013) como referência, dará auxílio ao sistema de indexação dos principais temas e, trará benefícios para a recuperação da informação. A estruturação de um esquema de classificação pode ser facilitada pela utilização de uma codificação numérica para designar as classes, subclasses, grupos e subgrupos preestabelecidos, o que agiliza a ordenação, a escolha do método de arquivamento e a localização, física e lógica (BRASIL, 2020, p. 7).

A partir desse ponto, a indexação será o elo entre a classificação de documentos e os principais temas, cuja finalidade será a recuperação do conteúdo das informações registradas no Fala.BR. Como aponta Cunha e Cavalcanti (2008, p. 193) “a indexação representa o conteúdo temático de um documento, por meio dos elementos de uma linguagem documentária, que indicam as palavras-chave ou descritores que representam os temas tratados num texto”. E segundo Strehl (1998 p. 329), a indexação assegura a recuperação de qualquer documento ou informação por meio da representação do seu assunto em sistemas de informação.

A relevância da indexação para os pedidos de acesso às informações é criar um instrumento facilitador de busca dos pedidos e respostas dos órgãos para os usuários, em consonância com a LAI, por meio de palavras-chave do Fala.BR, e ter mais acesso à informação. No caso em questão, o instrumento de referência para indexação dos Assuntos, Subassuntos e *Tags* é o Vocabulário Controlado do Governo Eletrônico (VCGE). Porém, este só contempla o Assunto, ficando a cargo do próprio usuário a indexação do Subassunto e da *Tag* (termos livres). O outro instrumento é o CCDA que fornece uma estrutura de classificação com Classes, Subclasses, Grupos, Subgrupos, como se observa na figura 2.

AKR Advances on Knowledge Representation Journal

Figura 2 - Estruturação de um esquema de classificação do Fala.BR e CCDA

Fonte: Santos Júnior; Corrêa; De Faria (2023)

A partir deste ponto, verifica-se a possibilidade de substituir os Assuntos, Subassuntos e *Tags* por Classes, Subclasses e Grupos na utilização do ChatGPT, o que poderá facilitar a busca da informação, dando aos usuários mais opções e recursos na escolha dos Assuntos e no preenchimento dos Subassuntos e *Tags* do Fala.BR.

2.3 ChatGPT

O ChatGPT é um sistema de conversação desenvolvido pela empresa OpenAI que utiliza a tecnologia de modelos de linguagem, especificamente o GPT (*Generative Pre-trained Transformer*), para interagir com os usuários em linguagem natural (Radford et al, 2018).

Criado no final de 2022, se tornou rapidamente uma importante ferramenta para criação e edição de texto, além de ser utilizado para consultas, com as devidas ressalvas, além de leitura de imagens e grandes volumes de dados (Lemos, 2023). Seguindo a evolução da arquitetura *Transformer*[5] para ChatGPT, Kocoń et al (2023), apresentam todas as versões desde o início, passando pelo ChatGPT até o estágio atual GPT-4: (1) modelo básico; (2) primeira versão do modelo de Pré-Treinamento Generativo (GPT); (3) GPT-2; (4) GPT-3; (5) *InstructGPT* baseado em *feedback* humano; (6) ChatGPT — um modelo que interage de forma conversacional, treinado em *feedback* mais humano; (7) GPT-4 — um modelo multimodal em grande escala com texto e/ou imagem como entrada.

De acordo com Deng e Lin (2022), o ChatGPT é um sistema de PNL (Processamento de Linguagem Natural) poderoso que pode gerar conversas semelhantes às humanas. Ele tem vários benefícios, como maior eficiência, maior precisão e economia de custos (comparado com o trabalho de um ser humano). No entanto, também apresenta alguns desafios, como preocupações de segurança e capacidades limitadas. Apesar destes

[5] É uma arquitetura de rede simples, baseada apenas em mecanismos de atenção, dispensando totalmente a recorrência e as convoluções (Vaswani, 2017).

AKR Advances on Knowledge Representation Journal

desafios, ChatGPT é uma tecnologia de Inteligência Artificial (IA) promissora que pode ser usada para automatizar conversas e gerar respostas mais precisas.

Outro modelo de processamento de linguagem natural é o BERT ou Bidirectional Encoder Representations from Transformers (Devlin et al. 2018), desenvolvido por pesquisadores do Google baseado em *transformers,* que são arquiteturas de redes neurais artificiais baseadas em atenção que consideram o contexto de todas as palavras no texto simultaneamente. O BERTimbau (Souza; Nogueira; Lotufo; 2020) é uma adaptação do BERT para o português brasileiro, desenvolvido com o objetivo de capturar melhor as especificidades linguísticas e culturais regionais, foi treinado para entender melhor a sintaxe e o uso de palavras no contexto brasileiro. Este modelo especializado resulta em uma melhoria significativa em várias tarefas de PLN em comparação com modelos multilíngues ou genéricos. Nesta pesquisa, o BERTimbau foi empregado para gerar os *word embeddings*[6] dos pedidos extraídos da plataforma Fala.BR.

3 Procedimentos Metodológicos

Trata-se de uma pesquisa exploratória e explicativa, de natureza aplicada, cuja abordagem é qualitativa-quantitativa e os procedimentos técnicos a serem utilizados serão o bibliográfico e experimental.

Para realização da pesquisa cumpriram-se 8 etapas: a) levantamento de informações, por meio do *Download* de Dados do Fala.BR de todos os pedidos de acesso à informação cadastrados no ano de 2023; b) seleção dos órgãos referentes às Universidades Federais, excluindo os Hospitais Universitários, CEFETs e Institutos Federais; c) separação apenas dos pedidos que contenham a palavra EDUCAÇÃO no campo "ResumoSolicitacao"; d) sorteio de 100 pedidos de acesso à informação através do método "*sample*" do pacote pandas da linguagem de programação *Python*[7]; e) utilização dos critérios de exclusão dos textos: abaixo de 50 e acima de 250 palavras, ou seja, somente aqueles que estão incluídos nessa faixa de palavras. A justificativa é que havendo poucas palavras de texto, pode ocasionar ausência na classificação ou rotulação de dados[8], bem como, o excesso de palavras poderá tornar a leitura dos especialistas mais

[6] É um componente essencial no processamento de linguagem natural (NLP) e de chatbots que desempenham um papel fundamental na compreensão e geração de respostas em conversas com usuários (Dendrites, 2024).

[7] Python é uma ótima linguagem para fazer análise de dados, principalmente por causa do fantástico ecossistema de pacotes python centrados em dados. O Pandas sample é usado para gerar uma linha ou coluna aleatória de amostra do quadro de dados do chamador da função (Acervo Lima, 2022).

[8] é um processo fundamental no campo da inteligência artificial e aprendizado de máquina, conhecida como anotação de dados, essa prática consiste em atribuir rótulos ou tags a um conjunto de dados para

AKR Advances on Knowledge Representation Journal

cansativa; f) convidar 10 (dez) especialistas (arquivistas) para produção de textos de "busca de pedidos e respostas" registrados no ano de 2023; g) arquivistas farão a classificação e indexação dos textos dos pedidos de acesso à informação do Fala.BR, utilizando como referência técnica o Código de Classificação de Documentos de Arquivo para atividades-fim da IFES; h) comparar os resultados dos materiais produzidos pelos arquivistas com o do ChatGPT a fim de avaliar o grau de eficiência da ferramenta.

De acordo com Kocoń et al (2023) existem três estágios do processamento de dados: (1) seleção de um conjunto de dados e conversão do conjunto de teste em formato baseado em *prompt*; (2) consultar (solicitar) o serviço ChatGPT usando a API da OpenAI; (3) extrair rótulos de resultados brutos e avaliar usando dados reais e comparar as respostas com os especialistas da informação (Arquivistas).

Para a geração do *prompt* (estágio 1), utilizou-se as informações do VCGE e do CCDA das Atividades-fim das IFES, no qual, foram adotados os seguintes critérios para os Assuntos: i) Educação básica; ii) Educação profissionalizante; iii) Educação superior; iv) Outros em Educação; v) sem relação com Educação para os Assuntos. O Subassunto está relacionado a um Assunto e que, obrigatoriamente, apareçam no Código de Classificação das IFES como uma subcategoria da CLASSE, seguindo uma taxonomia conforme o assunto. Já as TAGs, foram elencadas 5 (cinco) palavras-chave que representem bem o texto e que possam auxiliar na indexação do texto.

O estágio 2 envolve a utilização de uma API[9] para acessar o modelo de linguagem do ChatGPT, que está hospedado na nuvem. O *prompt* é inserido junto à *query*, constituída por pedidos de acesso à informação, e o modelo então executa a tarefa e retorna os resultados. Isso levanta uma questão de pesquisa: os grandes modelos de linguagem, empregando um "aprendizado em contexto"[10], são eficazes em gerar palavras-chave relevantes e classificar corretamente os principais temas em pedidos feitos na plataforma Fala.BR na área de Educação, usando apenas um vocabulário controlado? Para investigar essa questão, realizou-se um teste experimental com uma variedade de pedidos, explorando as capacidades do modelo.

Na etapa final do projeto (etapa 3), dez arquivistas participaram da pesquisa. Cada um recebeu cinco textos oriundos dos pedidos de acesso à informação para realizar análise, classificação e indexação dos Assuntos, Subassuntos e *Tags,* conforme

que eles possam ser corretamente interpretados e utilizados por algoritmos de machine learning (Dendrites, 2024).

[9] é uma tecnologia que permite a interação entre diferentes sistemas de software e a utilização de recursos de Inteligência Artificial (IA) em aplicações (Dendrites, 2024).

[10] é uma abordagem inovadora no campo da educação e do treinamento que combina elementos de diferentes disciplinas para criar um ambiente de aprendizado eficaz e personalizado (Dendrites, 2024).

AKR Advances on Knowledge Representation Journal

estabelecido pela plataforma Fala.BR, utilizando as mesmas informações do *prompt* da etapa 1 como material de referência. Após a obtenção dos dados rotulados pelos arquivistas e as classificações realizadas pelo modelo ChatGPT, os resultados foram avaliados por meio de métricas específicas. Para o campo "assunto", utilizou-se a acurácia e o teste do coeficiente Kappa de Cohen (Cohen, 1960), a fim de medir a concordância entre o especialista (arquivista) e o modelo (ChatGPT). A acurácia é uma métrica de desempenho que mede a proporção de classificações corretas entre todas as classificações realizadas. É definida como:

$$\text{Acurácia} = \frac{\text{Número de classificações corretas}}{\text{Número total de classificações}}$$

Se tivermos 100 classificações e 90 delas estão corretas, a acurácia seria:

$$\text{Acurácia} = \frac{90}{100} = 0.9 \text{ ou } 90\%$$

O coeficiente Kappa de Cohen é uma métrica estatística utilizada para medir a concordância entre dois avaliadores ou classificadores categóricos, levando em consideração a concordância que poderia ocorrer por acaso. Este coeficiente é útil em diversos campos, incluindo medicina, ciências sociais e análise de conteúdo, onde é importante avaliar a consistência entre diferentes avaliadores.

Definição: Seja P_0 a proporção de observações em que os avaliadores concordam e P_e a proporção de concordância esperada por acaso. O coeficiente Kappa (K) é calculado como:

$$K = P_0 - P_e \,/\, 1 - P_e$$

P_o: Proporção de concordância observada.

P_e: Proporção de concordância esperada por acaso.

Interpretação dos Valores de Kappa (K)

$K < 0$: Concordância pior do que o acaso $0.41 \leq K \leq 0.60$: Concordância moderada

$K = 0$: Concordância equivalente ao acaso $0.61 \leq K \leq 0.80$: Concordância substancial

$0 < K \leq 0.20$: Concordância ligeira $0.81 \leq K \leq 1.00$: Concordância quase perfeita

$0.21 \leq K \leq 0.40$: Concordância regular

Para os demais campos, como Subassuntos e *Tags*, foi utilizada uma representação do texto em espaço vetorial pelo modelo de linguagem BERTimbau (Souza; Nogueira; Lotufo; 2020), calculando a distância de cosseno para determinar a similaridade via Word Embedding, onde temos para os campos Subassuntos: valores altos de similaridade (próximo de 1), valores moderados de similaridade (0.5 a 0.7) e valores

AKR Advances on Knowledge Representation Journal

baixos de similaridade (abaixo de 0.5). E para as *Tags*: Valores altos de similaridade (próximo de 1), valores moderados de similaridade (0.6 a 0.8) e valores baixos de similaridade (abaixo de 0.6).

Concordância de *Tags*: A similaridade de cosseno para *Tags* é geralmente mais alta do que para subassuntos, sugerindo que o ChatGPT é mais eficaz em capturar e replicar as *Tags* associadas aos assuntos.

4 Análise e Resultados

4.1 Pedidos de Acesso à Informação das Universidades Federais do Brasil 2023

Os dados foram extraídos em março de 2024, por meio do *Download* de Dados, contendo todos os pedidos de acesso à informação das Universidades Federais do Brasil registrados entre 01 de janeiro de 2023 e 31 de dezembro de 2023, do Fala.BR. Obteve-se 90.183 dados dos pedidos de acesso à informação referentes a 323 órgãos. Destes, 11.941 pedidos eram exclusivamente das Universidades Federais. Na Tabela 1, tem-se uma síntese do levantamento dos Assuntos, Subassuntos e *Tags*.

Tabela 1 – Assuntos, Subassuntos e *Tags* das Universidades Federais em 2023

Descrição	ASSUNTOS	SUBASSUNTOS	*TAGS*
Cadastro	Obrigatório	Opcional	Opcional
Quantidade de registros	127 registrados	715 registrados	203 registrados
Quantidade de Universidades Federais	64 Universidades preencheram o campo	35 Universidades que preencheram o campo	33 Universidades que preencheram o campo
Termo mais registrado	Acesso à Informação (45,37%)	Transparência/ Institucional (0,69%)	SIC (1,30%)
Percentual de preenchimento	100%	18,85%	18,49%
Percentual de campo vazio	0%	81,15%	81,51%

Fonte: Dados da pesquisa

O campo Assunto "Acesso à Informação" teve 5.418 registros, equivalente a 45,37% dentre os 11.941 pedidos de acesso à informação registrados em 2023. Pressupõe-se que os demais 177 cadastrados não estão sendo bem aproveitados ou não são necessários para

AKR Advances on Knowledge Representation Journal

os usuários do Fala.BR. Ao contrário do Assunto, que tem uma lista fechada, os Subassuntos e as *Tags* são abertas e ficam a cargo do próprio usuário inserir a palavra-chave ou termo, o que demonstra maior quantidade de registros. Porém, há maior incidência de campos vazios. Um dos motivos se deve pelo fato do Assunto ser obrigatório e os demais opcionais. Além disso, não há critérios de padronização e o preenchimento dos Subassuntos e das *Tags* ficam a cargo do próprio usuário.

4.2 Uso do ChatGPT e de Arquivistas nos pedidos de acesso à informação do Fala.BR

Em abril de 2024 foram selecionados 100 textos dos pedidos de acesso à informação da "Busca de pedidos e respostas", cujo termo constasse a palavra EDUCAÇÃO. Em seguida, para a geração do *prompt* do ChatGPT, utilizaram-se as informações do VCGE e do CCDA das atividades-fim das IFES.

Após a análise dos 100 textos dos pedidos de acesso à informação pelo ChatGPT, adotou-se os critérios de exclusão (28 textos) e, dentre os 72 restantes, foram sorteados 50 textos e entregues para dez arquivistas (cinco para cada) que, da mesma forma do ChatGPT, utilizaram o CCDA das IFES e, em ambos os casos, tiveram todos os campos dos principais temas preenchidos. Na tabela 2 tem a análise dos resultados de conteúdo entre os Arquivistas e o ChatGPT.

Tabela 2 – Resultados da análise de conteúdo: Homem x Máquina

Principais Temas	**Análise de Concordância/ Similaridade de cosseno**	**Resultados**	**Média de concordância**
Assunto	Acurácia Kappa de Cohen	38% 0.1066	Concordância Ligeira
Subassunto	*BERTimbau e Word Embedding*	0.5945	Concordância Moderada
Tag	*BERTimbau e Word Embedding*	0.7702	Concordância Relativamente Alta

Fonte: Dados da pesquisa

Após a análise dos arquivistas, somando com os resultados do ChatGPT, resolveu-se avaliar o grau de efetividade dos resultados entre o homem x máquina. A acurácia para o campo 'ASSUNTO' foi de 0.38, indicando que 38% das classificações feitas pelo ChatGPT coincidiram exatamente com as do especialista humano, enquanto no coeficiente Kappa obteve-se um valor de 0.1066, indicando uma concordância ligeira para o campo Assunto. A similaridade dos valores de cosseno (BERTimbau) dos Subassuntos e das *Tags* apresentaram, aproximadamente, 0.5945 para os Subassuntos, o que sugere uma concordância moderada. E para as *Tag*s, aproximadamente 0.7702, o que sugere uma concordância relativamente alta.

AKR Advances on Knowledge Representation Journal

Para finalizar, o quadro 1 apresenta um exemplo de como se chegaram aos resultados dos principais temas, separados por temas e divididos pelos usuários que realizaram os registros em 2023, com a análise do ChatGPT e Arquivistas. Os erros de português (ortográficos ou gramaticais) que se encontram no texto não foram corrigidos, pois manteve-se a originalidade dos pedidos de acesso à informação do Fala.BR.

Quadro 1 – Resultado dos principais temas indexados

TEXTO 12	**Participação em pesquisa para Pós-graduação em ensino de matemática Pedido 23546079412202313 - 08/09/2023 - Acesso Concedido**		
Detalhamento Solicitacao	Prezados, Meu nome é [redacted], sou doutoranda do curso de Pós-graduação em Educação Matemática e Tecnológica da Universidade Federal de Pernambuco - EDUMATEC - UFPE. Estou desenvolvendo um projeto de pesquisa sobre questões étnico-raciais em cursos de pós-graduação em Educação Matemática. Um dos objetivos do meu trabalho é coletar informações acerca das ações desenvolvidas pelas instituições de Ensino Superior para promoção do enfrentamento ao racismo, em cursos de pós-graduação stricto sensu na área de ensino de matemática. Assim, solicito, gentilmente, apoio para minha pesquisa, respondendo às questões no corpo desse e-mail, pois sem essa participação não será possível identificar os dados iniciais para o desenvolvimento do meu estudo. Seguem as perguntas: 1. O curso dispõe de políticas públicas de ações-afirmativas para estudantes/professores negros? Quais? 2. No caso de acesso por cotas raciais, quantos estudantes acessaram o curso através da política de cotas em 2023? 3.O curso desenvolve grupos/linhas de pesquisa, disciplinas, eventos, formações, pesquisas que contemplem temáticas que promovam a equidade racial? Se sim, quais? Desde já sou imensamente grata pela participação. Atenciosamente,		
TEMA	**USUÁRIO FALA.BR** (Pedido Original)	**CHATGPT** (Análise Digital)	**ARQUIVISTA** (Análise profissional)
ASSUNTO	Acesso à informação	Ensino Superior	Pesquisa
SUBASSUNTO	-	Cursos de pós-graduação stricto sensu (inclusive na modalidade a distância)	Outros assuntos relacionados à pesquisa
TAG 1	-	Políticas públicas de ações-afirmativas	Pedido de informações
TAG 2	-	Cotas raciais	Ações afirmativas
TAG 3	-	Equidade racial	Cotas raciais
TAG 4	-	Grupos de pesquisa	Equidade racial
TAG 5	-	Eventos acadêmicos	Educação matemática

Fonte: Dados da pesquisa

Percebe-se que o usuário classificou apenas o Assunto, enquanto o ChatGPT e os Arquivistas preencheram todos os campos. Comparando os resultados entre os dois

AKR Advances on Knowledge Representation Journal

últimos, houve divergência entre a máquina e o humano no campo Assunto (Ensino Superior x Pesquisa). A análise do campo Subassunto entre o ChatGPT e o Arquivista é compatível (em parte), obteve-se a média de similaridade de cosseno em 0,495451. Entretanto, ao se fazer uma análise das *Tags,* o resultado foi satisfatório, chegando a média da similaridade de cosseno de 0.867704.

5 Considerações Finais

A pesquisa apontou que um único assunto "acesso à informação" possui quase a metade dos registros no preenchimento dos Subassuntos e *Tags,* além de uma grande ausência no preenchimento, superando os 80%. Dentre os principais motivos estão a enorme quantidade de assuntos listados (178), a falta de conhecimento, treinamento ou material de apoio acessível e atualizado para os usuários. A experiência utilizada nesta pesquisa baseou-se em informações do VCGE e do CCDA das IFES para alimentação do *prompt* para que pudesse fazer a leitura dos dados. Pode-se observar a inversão de valores no Fala.BR, pois o que era negativo (ausência de preenchimento em 80%) passou a ser positivo com uso do ChatGPT e dos Arquivistas (77% das *Tags* preenchidas).

A pesquisa demonstrou ainda que é possível trazer melhorias para o registro dos principais temas dos pedidos de acesso à informação utilizando o ChatGPT. Quanto às limitações encontradas, poderiam ser aperfeiçoadas se: houvessem mais especialistas para cada texto, tornando mais precisas as análises de Kappa de Cohen; ou treinar o GPT para agir como os especialistas, que usaram de sua experiência das atividades-meio do CCDA para classificar os textos; e ainda se tivesse a presença de bibliotecários na análise dos textos por meio do VCGE, proporcionando outros resultados e contribuições para o ChatGPT.

Quanto às propostas futuras, o ChatGPT pode vir a apresentar as opções de preenchimento para os usuários, inclusive, os Subassuntos e *Tags*; ou sugerir a inclusão do código de classificação de documentos de arquivo das atividades-meio como gestão de pessoas, materiais, financeira, entre outros; e também aproveitar a solução nas manifestações de Ouvidoria que utiliza o mesmo sistema.

Espera-se que a pesquisa traga reflexões, debates e aperfeiçoamento ao Fala.BR e interesse aos demais interessados, cuja proposta é beneficiar o cidadão no acesso à informação com mais agilidade, qualidade e transparência.

REFERÊNCIAS

Acervo Lima. Python | Pandas *Dataframe.Sample* (2022). Disponível em: https://acervolima.com/python-pandas-dataframe-sample/ Acesso em: 29 jun. 2024

AKR Advances on Knowledge Representation Journal

Brasil. (2011). Lei nº 12.527, de 18 de novembro de 2011. Regula o acesso a informações previsto no inciso XXXIII do art. 5o, no inciso II do § 3o do art. 37 e no § 2o do art. 216 da Constituição Federal; altera a Lei no 8.112, de 11 de dezembro de 1990; revoga a Lei no 11.111, de 5 de maio de 2005, e dispositivos da Lei no 8.159, de 8 de janeiro de 1991; e dá outras providências.

Brasil. (2023e). Controladoria-Geral da União. Secretaria de Transparência e Prevenção da Corrupção. Acesso à Informação. *Busca de Pedidos e Respostas.* Disponível em: https://buscalai.cgu.gov.br/ Acesso em: 21 nov. 2023.

Brasil. (2023d). Controladoria-Geral da União. Secretaria de Transparência e Prevenção da Corrupção. Acesso à Informação. Busca de Pedidos e Respostas. *Download de Dados.* Disponível em: https://buscalai.cgu.gov.br/DownloadDados/DownloadDados Acesso: 21 nov. 2023.

Brasil. (2023b). Controladoria-Geral da União. Secretaria de Transparência e Prevenção da Corrupção. *Fala.BR - Manual.* Disponível em: https://wiki.cgu.gov.br/index.php/Fala.BR_-_Manual Acesso em: 21 nov. 2023

Brasil. (2023c). Controladoria-Geral da União. *Painel LAI.* Disponível em: https://www.gov.br/acessoainformacao/pt-br/perguntas-frequentes/painel-lei-de-acesso-a-informacao Acesso em 11 out. 2023

Brasil. (2023a). Ministério da Gestão e da Inovação em Serviços Públicos. Governo Digital. Linha do Tempo. *Do Eletrônico ao Digital.*

Brasil. (2013). Ministério da Gestão e da Inovação em Serviços Públicos. Conselho Nacional de Arquivos. Portaria MEC Nº 1.224, de 18 de dezembro de 2013. *Código de Classificação de Documentos de Arquivo Relativos às Atividades-fim das Instituições Federais de Ensino Superior e na Tabela de Temporalidade e Destinação de Documentos de Arquivo Relativos às Atividades-Fim das Instituições Federais de Ensino Superior.* Rio de Janeiro: Arquivo Nacional.

Brasil. (2020). Ministério da Gestão e da Inovação em Serviços Públicos. Conselho Nacional de Arquivos. *Código de classificação e tabela de temporalidade e destinação de documentos relativos às atividades-meio do Poder Executivo Federal.* Rio de Janeiro: Arquivo Nacional.

Brasil. (2005). Ministério da Justiça. Arquivo Nacional. *Dicionário brasileiro de terminologia arquivística*. Rio de Janeiro: Arquivo Nacional, 232p.

Brasil. (2016). Ministério do Planejamento, Orçamento e Gestão. Secretaria de Logística e Tecnologia da Informação. *VCGE: Vocabulário de Governo Eletrônico* - Brasília: MP, SLTI, 2014. Versão: 2016.

Cavalcanti, C. R. (1978). *Indexação & tesauro; metodologia & técnicas: edição preliminar*. ABDF.

AKR Advances on Knowledge Representation Journal

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1), 37-46.

Dendrites.Io. (2024). *Glossário de Inteligência Artificial de A a Z*. Disponível em: https://dendrites.io/glossario-de-inteligencia-artificial/ Acesso: em 12 jul. 2024.

Deng, J., & Lin, Y. (2022). The benefits and challenges of ChatGPT: An overview. *Frontiers in Computing and Intelligent Systems*, 2(2), 81-83.

Devlin, J. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Do Nascimento Silva, M. et al. (2020). Classificação Documental: Relato de Experiência no Arquivo Central da Reitoria do Instituto Federal de Educação, Ciência e Tecnologia de Sergipe (IFS). *Biblionline*, João Pessoa, v. 16, n. 2, p. 41-54.

Kocoń, Jan et al. (2023). ChatGPT: Jack of all trades, master of none. *Information Fusion*, 99, 101861.

Lemos, Amanda. (2023). Inteligência artificial. O que é ChatGPT? Tudo que você precisa saber para usar a IA. *Exame.* Publicado em: 27 jul. 2023.

Lima, G. Â. D. (2021). Gênesis da classificação: uma análise de conteúdo a partir da definição. *Perspectivas em Ciência da Informação*, 26(01), 197-237.

Radford, A. (2018). *Improving language understanding by generative pre-training*.

Santos Júnior, Zenóbio; Corrêa, Fábio. (2022). FALA.BR: entraves para gestão da ouvidoria no âmbito das manifestações designadas as Instituições Federais de Ensino Superior. *Analisando em Ciência da Informação*: Revista do Centro de Ciências Biológicas e Sociais Aplicadas (CCBSA) da Universidade Estadual da Paraíba (UEPB). v. 10, n. 1 (jan./jun. 2022) - João Pessoa: UEPB, p. 13-24.

Santos Júnior, Zenóbio; Corrêa, Fábio; De Faria, Vinícius F. (2023). Classificação de documentos de arquivo: principais temas dos pedidos de acesso à informação nas universidades federais mais demandadas pela LAI. *Arquivos, democracia e justiça social [livro eletrônico]* / organização Mariana Lousada Pinha, Marcia Cristina de Carvalho Pazin Vitoriano, Paulo Roberto Elian dos Santos. -- 1. ed. -- São Paulo: ARQ-SP,

De Sousa, R. T. B., & Junior, R. H. D. A. (2017). A classificação e o vocabulário controlado como instrumentos efetivos para a recuperação da informação arquivística. *In: Nelson Vaquinhas; Marisa Caixas; Helena Vinagre.(Org.). Da produção à preservação informacional: desafios e oportunidades*.

Souza, R. F. D. (2000). *A classificação como interface da Internet*.

AKR *Advances on Knowledge Representation Journal*

Souza, F., Nogueira, R., & 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 (pp. 403-417). Springer International Publishing.

Strehl, L. (1998). Avaliação da consistência da indexação realizada em uma biblioteca universitária de artes. *Ciência da Informação*, 27, 329-335.

Vaswani, A. et al. (2017). Attention is all you need. Advances in neural information processing systems. *Advances in neural information processing systems*, 30.