Universidade de Brasília – UnB

Faculdade de Ciência da Informação – FCI

Arquivologia

Inteligência artificial para classificação de documentos

Autora: Thalita Alves Cabral

Orientador: Prof. Dr. Renato Tarciso Barbosa de Sousa

Brasília, DF

2024

Thalita Alves Cabral

Inteligência artificial para classificação de documentos

Trabalho de Conclusão de Curso submetido ao curso de graduação em Arquivologia da Universidade de Brasília, como requisito para obtenção do Título de Bacharel em Arquivologia.

Universidade de Brasília – UnB

Faculdade de Ciência da Informação – FCI

Orientador: Prof. Dr. Renato Tarciso Barbosa de Sousa

Brasília, DF

2024

Universidade de Brasília

Faculdade de CIÊNCIA DA INFORMAÇÃO

THALITA ALVES CABRAL

INTELIGÊNCIA ARTIFICIAL PARA CLASSIFICAÇÃO DE DOCUMENTOS.

Monografia submetida ao corpo docente do Curso de Graduação em Arquivologia, da Faculdade de Ciência da Informação da Universidade de Brasília – UnB, como parte dos requisitos necessários à obtenção do título de Bacharel em Arquivologia.

Aprovado por:

Renato Tarciso Barbosa de Sousa	Rogério Henrique de Araújo Junior	Paulo José Viana de Alencar
Professor do Magistério Superior	Professor do Magistério Superior	Professor do Magistério Superior
Doutor	Doutor	Doutor

Documento assinado eletronicamente por **Renato Tarciso Barbosa de Sousa**, **Professor(a) de Magistério Superior da Faculdade de Ciência da Informação**, em 27/09/2024, às 08:16, conforme horário oficial de Brasília, com fundamento na Instrução da Reitoria 0003/2016 da Universidade de Brasília.

A autenticidade deste documento pode ser conferida no site http://sei.unb.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **11759432** e o código CRC **064CCA15**.

Olhai para as aves do céu, que não semeiam, nem ceifam,
nem ajuntam em celeiros; e vosso Pai celestial as alimenta.
Não valeis vós muito mais do que elas?
(Mateus 6:26)

Agradecimentos

A conclusão deste trabalho reflete ao fim de um ciclo muito importante em minha vida acadêmica, o caminho até aqui foi cheio de aprendizado e superação, o apoio de diversas pessoas fez a diferença contribuindo para esta jornada. Primeiramente, expresso minha gratidão a Deus, pela força que me deu em todos os momentos e por ter me guiado em todas as fases deste processo, me dando impulso para seguir em frente, fazendo com que cada obstáculo fosse uma oportunidade de crescimento e cada etapa alcançada um sinônimo de felicidade.

À minha família, que sempre esteve comigo e ofereceu suporte, incentivo e amor. Dedico meu agradecimento à minha mãe e irmãos, vocês foram combustível que me impulsionou a não desistir, por acreditarem em mim, minha sincera gratidão. Mãe, em cada gesto e palavra sinto o seu amor, sua coragem para enfrentar os desafios sempre foram, para mim, exemplos constantes de força. Aos meus irmãos, minhas inspirações desde pequena, agradeço por estarem ao meu lado, me apoiando de maneira sempre divertida, fazendo o caminho mais leve.

Ao meu orientador, Renato Tarciso Barbosa de Sousa, manifesto minha gratidão pela paciência, dedicação e conselhos que guiaram para o desenvolvimento deste trabalho. Aos meus amigos de curso, sou grata pela troca de experiências, as discussões construtivas e o apoio mútuo que cooperaram para a construção dessa trajetória, vocês estiveram presentes nos dias longos de estudo e nos momentos de descontração. Sem vocês, este percurso teria sido ainda mais solitário e difícil.

À pessoa especial que esteve ao meu lado em todos os momentos, meu companheiro, meu sincero agradecimento. O seu amor, suporte e presença me apoiou a chegar até aqui. Obrigada pela participação propondo boas percepções que enriqueceram ainda mais a pesquisa. Por sempre me incentivar e me motivar, agradeço. Por fim, estendo minha gratidão a todas as pessoas que fizeram parte direta ou indiretamente dessa trajetória. Cada contribuição, teve grande valor para o término desta importante etapa.

Resumo

Este trabalho aborda a classificação de documentos de arquivo feita por meio da inteligência artificial (IA), a fim de mostrar que ela é uma alternativa viável e que pode ser utilizada nos órgãos e instituições. O objetivo é avaliar a viabilidade de realização da classificação arquivística utilizando inteligência artificial. A metodologia utilizada será a identificação da produção científica sobre a classificação funcional, avaliação da usabilidade da classificação funcional a partir da literatura e o desenvolvimento de uma prova de conceito sobre a classificação por meio da inteligência artificial. Caracteriza-se como uma pesquisa qualitativa, com uma abordagem reflexiva na discussão, e explicativa, utilizando técnicas de estudo de caso no Senado Federal. Resultados esperados, busca-se solucionar os problemas da classificação manual identificados atualmente, trazendo a IA para a classificação, podendo ser inserida nos sistemas informatizados de gestão de documentos.

Palavras-chave: classificação arquivística; classificação funcional; inteligência artificial.

Abstract

This work addresses the classification of archival records using artificial intelligence (AI), in order to show that it is a viable alternative that can be used in agencies and institutions. The goal is to evaluate the viability of archival classification using artificial intelligence. The methodology used will be the identification of scientific production on functional classification, evaluation of the usability of functional classification based on literature and the development of a proof of concept on classification through artificial intelligence. It is characterized as a qualitative research, with a reflective approach in the discussion, and explanatory, using case study techniques in the Federal Senate. The aim is to solve the functional classification problems currently identified, bringing AI to classification, which can be incorporated into record management systems.

Key-words: archival classification; functional classification; artificial intelligence.

Lista de Quadros

Lista de Figuras

Sumário

1 Introdução

Com o avanço da sociedade, por meio da intensificação das atividades desenvolvidas pelo Estado, ocorreu um aumento crescente e significativo da produção de registros documentais arquivísticos, que foram acumulados ao longo do tempo. Desta forma, é fundamental cuidar de uma série de aspectos da documentação, para se ter uma eficiência maior no armazenamento e no uso. É indispensável um desenvolvimento teórico, através do princípio da proveniência e do princípio da ordem original, que são fundamentais para a classificação na arquivologia, preservando a organicidade e mantendo juntos aqueles documentos que se originam da mesma função.

A humanidade sempre teve a necessidade de registrar suas atividades e por meio da arqueologia, descobre-se como os registros e meios informacionais das comunidades no passado foram organizados, seja de forma produzida ou recebida, e pela necessidade de consolidar as maneiras de recuperação e organização dos registros documentais "desde que se começou a registrar a história em documentos, surgiu para o homem o problema de organizá-los" (SCHELLENBERG, 2004, p. 97). Com isso, os documentos foram criados, preservados e classificados conforme a vida política, cultural, religiosa e social.

Este trabalho é voltado para a compreensão do desenvolvimento do conceito de classificação automática de documentos, a partir da inteligência artificial (IA), como alternativa à classificação manual. Busca-se apresentar a discussão sobre os problemas trazidos pela classificação funcional dos documentos de arquivo. Seguidamente, foram apresentadas as definições e o surgimento do conceito mediante um levantamento bibliográfico do que se tem até os dias atuais. O foco é voltado para as bases teóricas e metodológicas pesquisando os princípios e fundamentos.

A classificação parte de um processo de organização intelectual no qual as informações e o contexto em que cada documento está inserido refletem as funções e atividades desenvolvidas pela pessoa, órgão ou instituição que os acumulou. A autora Martín-Pozuelo Campillos Martín-Pozuelo Campillos (1996, p. 54-55, tradução nossa) descreve que "Seguindo os canadenses Couture e Rousseau e desde uma perspectiva puramente metodológica, entendo que a classificação é a primeira etapa de um tratamento que conduz à acessibilidade do acervo documental."[1] A autora relata uma visão das definições de classificação no século XIX, onde ela defende que a preocupação com a classificação não era colocada de forma muito aberta, mas que, ainda assim, abria espaço para discussão da terminologia dessa atividade. Essa terminologia estava muito conectada com a parte mecânica em que o arquivo se colocava para verificar as suas funções, que eram tanto a de difundir as informações quanto a de receber os documentos de instituições, dando-lhes o devido tratamento. É apresentado o princípio da proveniência e como, por meio dele, é possível definir algumas bases teóricas para o entendimento da classificação

[1] "Siguiendo a los canadienses Couture y Rousseau y desde una perspectiva puramente metodológica, entiendo que la clasificación es la primera etapa de un tratamiento que conduce a la accesibilidad del acervo documental."

arquivística.

Schellenberg (1980) identificou três fases para o arranjo de documentos. A primeira fase consistia na classificação dos documentos de forma semelhante à classificação de livros em bibliotecas, onde o princípio da proveniência não era considerado na formulação da classificação. Para essa fase, os métodos utilizados foram baseados em assunto, ordem cronológica e espaço geográfico, além do agrupamento por atividade ou tipo documental. Na segunda fase, o princípio da proveniência foi utilizado como base para a ordenação dos documentos públicos e coleções de manuscritos, sendo inicialmente aplicado aos papéis públicos da Pensilvânia. Por fim, uma terceira fase consistiu na extensão do princípio da proveniência, na qual os documentos foram arranjados de acordo com sua origem em uma atividade orgânica. Schellenberg em seus livros "Arquivos Modernos: Princípios e Técnicas"(2004) e "Documentos Públicos e Privados: Arranjo e Descrição"(1980) diferencia os termos "classificação"e "arranjo", pois para ele havia uma assimetria entre os arquivos correntes e permanentes, e o arquivista profissional deveria praticar atividades específicas, o autor aborda esses procedimentos técnicos para implementar a classificação nos arquivos, assim como também faz uma distinção entre o arranjo de documentos arquivísticos e a classificação utilizada em bibliotecas.

> Os documentos textuais dividem-se em inúmeros tipos, dos quais alguns são as cartas, os relatórios, as contas e os diários. Há que tratá-los de per si, ou por conjuntos de várias espécies. Podem ser reunidos em unidades de arquivamento de diversas categorias... As publicações, por outro lado, não são divisíveis em numerosos tipos físicos. Consistem, normalmente, de livros impressos, séries ou edições por processos substitutos da tipografia. Trata-se, em geral, de unidades distintas, embora os textos seriados, tais como anuários e periódicos, se aproximem, de algum modo, das séries de arquivo, na medida em que equivalem a conjuntos de itens. Os documentos textuais são, de costume, o produto de atividade orgânica e revelam-se, portanto, significativos, mormente em relação àquela que resultou na sua criação... As publicações, por outro lado, são o produto de atividade cultural e aparecem significativas, em primeiro lugar, em relação aos assuntos de que tratam. As diferenças físicas tornam necessário ordenar os documentos como unidades coletivas e as publicações como unidades singulares. Em virtude das suas diferenças substantivas, é mister se ordenem os documentos pela origem e as publicações pelo assunto. As diferenças substantivas e físicas entre publicações e documentos fazem do arranjo do arquivo e da classificação da biblioteca duas operações radicalmente distintas (SCHELLENBERG, 1980, p.90-92).

No tratamento documental, é indispensável aos órgãos que se preocupam com a guarda de documentos, encontrá-los de forma rápida quando for solicitado, logo é importante que os documentos sejam conservados de maneira ordenada e acessível. Para que isso ocorra, é necessário que os documentos sejam bem classificados e bem arquivados (SCHELLENBERG, 2004). A partir disso, se os documentos forem corretamente classificados atenderá bem as necessidades do órgão, refletindo a função e as atividades que integram o local. O autor prossegue referindo-se sobre o cumprimento das funções básicas de uma repartição pública, sendo duas principais, atividade fim que reflete o trabalho técnico e profissional do órgão, e a atividade

meio que relaciona com a administração interna da organização, o que é comum e básico a todos os órgãos. É analisada, na classificação de documentos, a organização da entidade criadora, visto que os documentos são agrupados de modo a refletir a estrutura orgânica da entidade e essa estrutura é estabelecida pelos objetivos ou funções que se destinou aos órgãos.

Para a elaboração de um sistema de classificação, Schellenberg (2004) apresenta três tipos de sistemas de classificação, sendo a classificação por assunto, a classificação estrutural e a classificação funcional. A classificação por assunto, de acordo com Schellenberg, foi descartada por apresentar instabilidades, como a amplitude dos assuntos, a subjetividade, a inviabilidade, a acomodação para lidar com a diversidade de conteúdo e o tratamento com grandes volumes documentais. O autor compreendia que fazer um plano de classificação fundamentado na divisão em classes organizacionais não era o melhor, visto que a estrutura organizacional dos órgãos do governo é muito contingente para que se pudesse fazer uma classificação determinada dos documentos e o último aspecto é baseado nas funções da organização, de modo que os documentos são originados de uma função e sua classificação deve segui-la.

A historiadora e arquivista espanhola Heredia Hererra (1991, p. 268, tradução nossa) discorre sobre dois níveis para a classificação "O primeiro identifica-se com a estrutura ou funcionamento da instituição (por exemplo, o seu organograma: órgãos ou funções) e corresponde às seções e subseções; o segundo nível equivale às séries documentais, isto é, aos testemunhos de atividades derivadas daquela estrutura."[2] A organização em um arquivo, nada mais é, do que traduzir em um fundo documental o estado original de sua produção e crescimento. Isso traz efeitos tanto para os documentos quanto para as informações, estabelecendo sua relação entre os documentos e contribuindo na sua localização. A autora destaca também que classificar e ordenar são duas operações dentro de uma mais ampla que podemos chamar de organização, perfeitamente diferenciadas e indispensáveis, para conservação dos documentos.

A classificação e organização passaram a ser consideradas como atividades distintas, algumas vezes identificadas com a disposição numérica dos arquivos, também confundida com a distribuição por assuntos e datas. Hoje em dia, a classificação de qualquer fundo deve ser baseada em um plano estruturado que reflita as funções e atividades da instituição de onde provém, com isso o arquivista deve criá-lo de forma a posteriori, baseando-se no princípio da proveniência e os agrupamentos documentais que eles derivam (HEREDIA HERERRA, 1991). Partindo dessa abordagem, na classificação as classes devem ser feitas mediante a necessidade e não a priori de forma previamente já criadas.

O plano de classificação de documentos de arquivo é um instrumento de trabalho usado para classificar todo documento produzido ou recebido por uma instituição no exercício de suas funções e atividades, com objetivo de agrupar documentos com uma mesma característica,

[2] "el primero se identifica con la estructura o funcionamiento de la institución (por ejemplo, su organigrama: Órganos o funciones) y corresponde a las secciones y subsecciones; el segundo nivel equivale a las series documentales, es decir a los testimonios de actividades derivadas de aquella estructura."

sendo pelo conteúdo e apresentá-lo por uma codificação que atinge até mesmo a organização física dos documentos no seu acondicionamento (RIOS; CORDEIRO, 2010). É possível a partir disso, colocar de forma visível as dificuldades que passam uma instituição, pois o instrumento permite ver todas as divisões da massa documental, isso traz um auxílio para a recuperação dos documentos no acervo.

Com o progresso dos novos comportamentos informacionais, a demanda por um instrumento arquivístico mais sofisticado deve ser requisitada, pois a consequência disso seria dar uma possibilidade maior de busca, representação e inclusão, isso reflete de forma importante no crescimento da classificação que é uma função arquivística. Contudo, a atividade fundamental de organização de documentos foi deixada de lado em vista de novos temas, com isso percebesse o motivo da baixa literatura sobre um aprimoramento da classificação de documentos, a pobreza de pesquisa traz um impacto importante nas formulações de sistemas automatizados, na implementação de instrumentos de classificação e na literatura (MOKHTAR; YUSOF, 2015).

A classificação funcional é um pilar da metodologia na arquivologia, o esquema de classificação funcional uniforme para os documentos de arquivo é um dos instrumentos mais importantes na gestão de documentos e informações nas organizações, esses esquemas são projetados para facilitar a criação e recuperação de registros, incluindo os documentos digitais, especialmente quando grandes quantidades de informações estão envolvidas (GUNNLAUGSDOTTIR, 2012). A ação de classificar os documentos de arquivo assegura os seguintes aspectos: a recuperação rápida e eficaz de informações, facilitando a busca e a localização dos documentos quando demandada, e a manutenção do vínculo arquivístico, onde a organicidade reflete a interconexão dos documentos dentro do contexto da organização. E por sua vez, a classificação fundamenta a avaliação e a descrição, facilitando a recuperação e compreensão do conteúdo dos documentos.

A estrutura deste trabalho abrange uma sessão que analisou como se constatou e efetuou a classificação na arquivologia até o surgimento da classificação funcional. Nas demais seções, é apresentado o enredo de inteligência artificial, avaliação da classificação funcional e por fim a classificação por meio da inteligência artificial.

1.1 Identificação do problema

A discussão que apresentamos aqui trata sobre a classificação de documentos de arquivo feita por meio da inteligência artificial. Seria essa uma solução para os problemas que são trazidos a partir da classificação feita manualmente pelos usuários?

1.2 Objetivo geral

Avaliar a viabilidade de realização da classificação arquivística utilizando inteligência artificial.

1.3 Objetivos específicos

- Identificar a produção científica sobre classificação funcional.
- Avaliar a usabilidade da classificação funcional a partir da literatura.
- Desenvolver uma prova de conceito sobre a classificação por meio da inteligência artificial.

1.4 Justificativa

Nos dias atuais, os arquivos são encontrados na forma híbrida, tendo uma quantidade alta de massa documental acumulada em suporte digital e papel, é gerada cada vez mais a criação de documentos digitais, com um superficial tratamento arquivístico, voltado para a correta classificação documental. De modo que, uma classificação incorreta torna um documento perdido, no qual não pode ser encontrado, resultando também em uma aplicação da tabela de temporalidade errada além de uma descrição insuficiente. A tecnologia vem avançando fortemente na arquivologia, a classificação de documentos não pode ser deixada para trás, visto que nas vertentes da gestão documental ela já está presente. Com o uso da IA, podemos obter resultados precisos e corretos de uma boa classificação documental, gerando sempre dentro do órgão ou instituição, o adequado tratamento e não mais uma classificação de forma subjetiva e por muitas vezes incorreta.

As vantagens do uso da IA para a classificação de documentos, está em uma maior agilidade para efetivação desta ação, poupando tempo, visto que, de outra forma seria realizada de maneira manual. Logo em seguida, podemos constatar que o grande volume documental pode ser tratado pela inteligência artificial, bem como manter o padrão de qualidade da operação, o que torna algo agradável às instituições com alta produção ou recebimento de documentos.

O modelo gerado pela IA pode ser moldado e atualizado para trabalhar com novas classificações de documentos, isso traz flexibilidade para que esteja sempre em concordância com as necessidades de mudança da instituição. É importante salientar também que a classificação correta, contém conformidade com a lei e os padrões e normas relacionadas à segurança da informação, onde documentos que precisam ser classificados com sigilo podem ser automaticamente protegidos pelo grau de sigilo vinculado à classificação, funcionalidade que pode ser criada no sistema informatizado de gestão arquivística de documentos, onde apenas pessoas que tenham delegação, poderão acessar as informações sensíveis contidas nos documentos. Portanto, a classificação correta, resposta que pode ser obtida de forma assertiva e objetiva mediante a inteligência artificial, oferece uma realidade essencial para o desempenho e sucesso do órgão ou instituição.

Foi feita uma pesquisa na literatura sobre o tema abordado neste trabalho e foram encontrados pouquíssimos textos publicados, isso gerou curiosidade e interesse para desenvolver

e pesquisar sobre a classificação arquivística de documentos realizada pela inteligência artificial. Desse modo, este texto representa uma valiosa contribuição para o avanço da área, visto que, ao desenvolver um tema pouco explorado e oferecer insights inovadores, é acrescentado mais conhecimento à literatura, fornecendo um panorama inédito e esclarecedor sobre o tema, tendo potencial de influenciar debates acadêmicos e direcionar práticas no campo. Sendo também um convite para futuros desenvolvimentos, evidenciando-se como uma referência benéfica para profissionais e estudiosos interessados no assunto.

1.5 Metodologia

A metodologia utilizada é uma pesquisa qualitativa, com uma abordagem reflexiva na discussão, e explicativa utilizando técnicas de estudo de caso no Senado Federal. Será identificado a produção científica sobre a classificação funcional, avaliação da usabilidade da classificação funcional a partir da literatura e o desenvolvimento de uma prova de conceito sobre a classificação por meio da inteligência artificial.

2 Referencial Teórico

Um marco para a história da classificação dos documentos arquivísticos foi a partir da declaração dos documentos da nação francesa que tornou como propriedade pública, abrindo o acesso para que os cidadãos possam consultar seus direitos individuais e coletivos, ela reconhece também a igualdade, especialmente perante a lei e a justiça, e isso determinou a propagação da legislação e regulamentações dirigidas a proteger o contexto documental, isso gerou a proclamação do princípio de acesso do público aos arquivos, e o reconhecimento da responsabilidade do Estado pela conservação dos documentos de valor, assim como também a criação de uma administração nacional e independente dos arquivos.

> A classificação é uma função importante para a transparência e o compartilhamento de informações, que são caminhos seguros para a tomada de decisão, para a preservação da memória técnica e administrativa das organizações contemporâneas e para o pleno exercício da cidadania. Ela é uma atividade reconhecida, pela maior parte dos autores que tratam da questão, como matricial. Ela precede todas as outras atividades (SOUSA, 2003, p. 240).

A classificação funcional é definida pelo Glossary of terms, de 2009, do International Records Management Trust, como um sistema para organizar materiais com base na função, atividade praticada por uma organização para cumprir seu mandato, e não apenas de fundamentado em sessão ou assunto. Contudo, para a classificação funcional, não existe um modelo comum que se baseia nas funções, visto que cada arquivo dentro da instituição ou órgão desempenha finalidades diferentes entre si, com isso a nomenclatura também não mantém um padrão (SOUSA, 2005).

A inteligência artificial é a parte da ciência da computação voltada para o desenvolvimento de sistemas ou máquinas que executem tarefas que necessitam da inteligência humana, podendo ser aprimorado de forma interativa. O aprendizado de uma máquina é desenvolvido por meio da experiência, conforme a atividade é executada.

> Um processo de aprendizagem inclui a aquisição de novas formas de conhecimento: o desenvolvimento motor e a habilidade cognitiva (através de instruções ou prática), a organização do novo conhecimento (representações efetivas) e as descobertas de novos fatos e teorias através da observação e experimentação. Desde o início da era dos computadores, têm sido realizadas pesquisas para implantar algumas destas capacidades em computadores. Resolver este problema tem sido o maior desafio para os pesquisadores de inteligência artificial (IA). O estudo e a modelagem de processos de aprendizagem em computadores e suas múltiplas manifestações constituem o objetivo principal do estudo de aprendizado de máquinas (SANTOS, 2005).

Desse modo, o aprendizado pode ser adquirido por hábito, onde o programa aprende pela prática, seguindo o que foi informado. Isso gera novas formas de programação, tendo cada vez mais desafios a serem ultrapassados, se tornando cada vez melhor. Bellotto (2006, p. 299) afirma o seguinte: "O arquivista hoje não pode esquecer que vive e atua profissionalmente na

chamada “era da informação”, na qual as tecnologias da informação e da comunicação tem presença marcante.” Desse modo, observa-se que a arquivologia assim como as outras áreas do conhecimento vem sendo atingida pelo avanço da tecnologia, isso afeta também os meios de fornecimento e armazenamento da informação. Por meio da informática, foi feita a criação de ferramentas que gerenciam a informação, além de novos suportes da informação.

O sistema informatizado de gestão arquivística de documentos (SIGAD) tem sua criação por meio da implementação de uma política arquivística no órgão ou entidade para organizar e gerenciar seus documentos de forma eficiente, garantindo sua acessibilidade, autenticidade e preservação ao longo do tempo. Ajudando a lidar com a gestão da informação, onde naturalmente se encontra um volume crescente de documentos.

> É um conjunto de procedimentos e operações técnicas, característico do sistema de gestão arquivística de documentos, processado por computador. Pode compreender um software particular, um determinado número de softwares integrados, adquiridos ou desenvolvidos por encomenda, ou uma combinação destes. O sucesso do SIGAD dependerá, fundamentalmente, da implementação prévia de um programa de gestão arquivística de documentos (E-ARQ BRASIL, 2011, p. 10).

No Quadro 1, abaixo, temos as diretrizes para a elaboração do projeto de pesquisa apresentado.

Quadro 1 – Elaboração do projeto de pesquisa

Tema	**Problema**	**Objetivo Geral**	**Objetivos específico**	**Metodologia**	**Fonte**
Inteligência artificial para classificação de documentos.	A discussão que apresentamos aqui trata sobre a classificação de documentos de arquivo feita por meio da inteligência artificial. Seria essa uma solução para os problemas que são trazidos a partir da classificação feita manualmente pelos usuários?	Avaliar a viabilidade de realização da classificação arquivística utilizando inteligência artificial.	Identificar a produção científica sobre classificação funcional.	Revisão de literatura sobre a classificação funcional.	Literatura da área.
			Avaliar a usabilidade da classificação funcional a partir da literatura.	Revisão de literatura sobre a usabilidade da classificação funcional de documentos de arquivo.	Literatura da área.
			Desenvolver uma prova de conceito sobre a classificação por meio da inteligência artificial.	Pesquisa qualitativa utilizando técnicas de estudo de caso.	Senado Federal.

Fonte: elaboração própria

3 Revisão de Literatura

3.1 Inteligência artificial

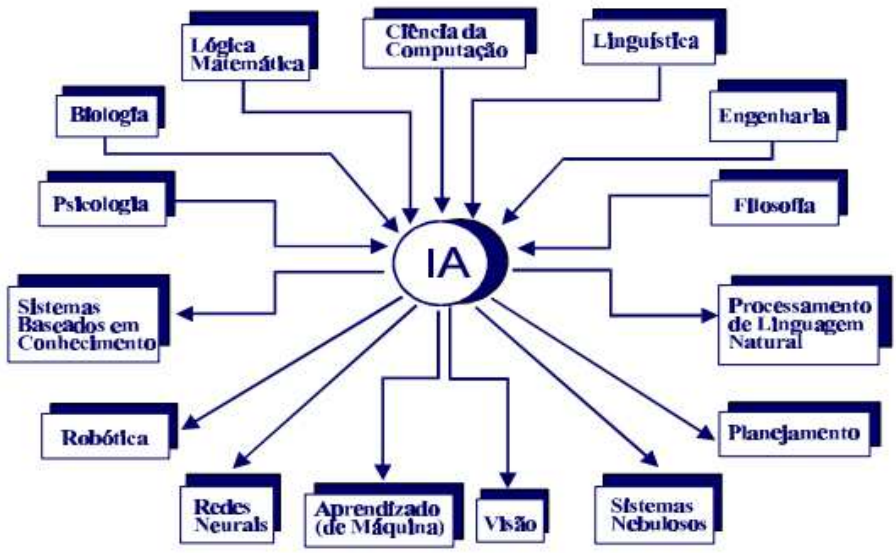
A mudança para o meio digital tem um impacto profundo na forma como são gerados os documentos, necessitando de uma alta infraestrutura tecnológica e de gestão para se adaptarem às novas condições do avanço tecnológico. A digitalização de documentos facilita a organização, armazenamento e recuperação de informações, contribuindo de forma rápida e eficiente nas tarefas administrativas. Estas soluções tecnológicas cooperam para diminuir o espaço físico necessário para armazenamento, reduzem o risco de perda ou deterioração de documentos e fornecem e protegem a informação de forma segura e acessível.

> O conhecimento quando preservado a partir dos documentos arquivísticos digitais e seus processos informatizados permitirá seu compartilhamento no futuro. Compartilhamento que garantirá a perpetuação de conhecimento explícito seja qual for à tecnologia utilizada pelas próximas gerações (INNARELLI, 2012, p. 62).

Inteligência Artificial (IA) provém da ciência da computação, Kurzweil (1990) fala sobre a arte de criar sistemas que possam executar atividades que exijam inteligência humana, o que resume bem a IA. Essas atividades comportam aprendizagem, raciocínio, habilidades de linguagem natural, identificar princípios e resolução de problemas. É uma tentativa de fazer com que a tecnologia digital aja como seres humanos. Isso atinge diversas áreas das ciências humanas, favorecendo o avanço em como os problemas são solucionados.

A inteligência artificial pode ser apresentada em diversas subcategorias, como o aprendizado de máquina, onde os computadores aprendem com os dados e melhoram o seu desempenho em determinadas atividades; visão computacional, que permite às máquinas interpretar e compreender imagens e vídeos; e processamento de linguagem natural, que engloba a interação entre computadores e a fala humana. Na sua essência, a IA procura reproduzir a inteligência dos humanos em equipamentos, permitindo-lhes raciocinar, aprender e adaptar-se. A IA tem muitas aplicações, oferecendo novas soluções para os problemas do homem e melhorando sua interação com a tecnologia, como mostrado na figura abaixo:

Figura 1 – Áreas relacionadas com a inteligência artificial



Fonte: (MONARD; BARANAUSKAS, 2000, p. 2)

Da Figura 1, destacam-se as seguintes aplicações: aprendizado de máquina, redes neurais e processamento de linguagem natural. O aprendizado de máquina, trata-se do desenvolvimento de programas com condição de executar uma atividade pela sua própria experiência (FACELI et al., 2011). Obtendo programas que aprendem sozinhos, utilizando um conjunto de dados que representam as experiências anteriores, um exemplo de tarefas de aprendizado de máquina, é a classificação e agrupamento de dados. Este processo integra probabilidade, IA, estatística, entre outras áreas de pesquisa.

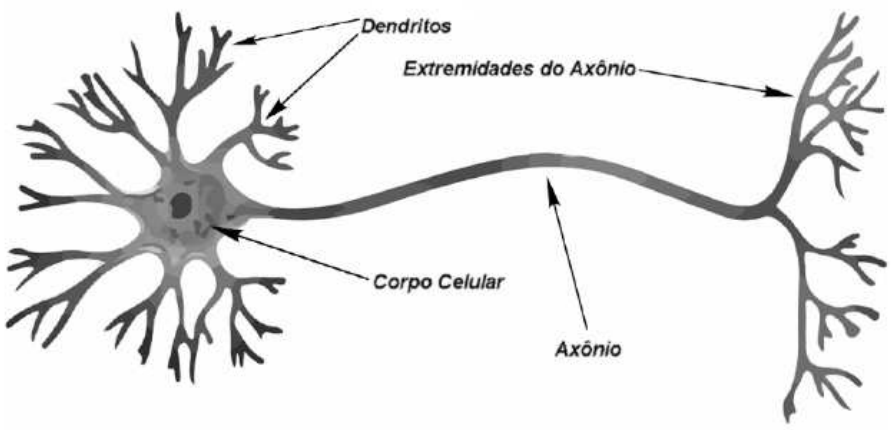
Avançando no aprendizado de máquina, encontra-se tarefas descritivas e preditivas. As tarefas descritivas, visam o progresso de algoritmos que descreverão os dados, como por exemplo o agrupamento de dados (HORTA; CAMPELLO, 2012) fazendo uma distinção para que os dados semelhantes fiquem no mesmo grupo, realizando uma busca por grupos onde o distanciamento entre seus integrantes seja o menor possível, ao mesmo tempo, distanciando grupos diferentes. Consequentemente, a distância entre dados de um mesmo grupo será bem menor do que a distância de dados de grupos distintos. O agrupamento de texto é um exemplo de aplicação, onde o algoritmo agrupa textos que apresentam o mesmo assunto e separa em grupos diferentes os textos que discutem assuntos dessemelhantes.

As tarefas preditivas podem ser separadas em classificação e regressão. Em classificação, procura-se atribuir o dado à categoria, o modelo aprende a atribuir entradas a uma ou mais classes tendo como suporte os exemplos que foram disponibilizados, como a classificação de texto (PHAN; NAKAGAWA, 2016). Já a regressão, é utilizada quando a saída prevê o valor de uma variável numérico contínuo, dadas outras variáveis (atributos de entrada), gerando uma função que esquematize os dados de entrada para um valor numérico, diminuindo a diferença entre as predições do modelo e os reais valores.

Redes neurais artificiais (RNA) são algoritmos que apresentam um modelo matemático baseado na estrutura biológica do sistema nervoso humano. A RNA corresponde a um esquema de processamento capaz de armazenar conhecimento fundamentado em aprendizagem e disponibilizar este conhecimento para a utilização a qual se destina (SPöRL; CASTRO; LUCHIARI, 2011). Por conseguinte, as redes neurais artificiais são capazes de aprender e tomar decisões que são observadas em seu próprio aprendizado.

Com base em Haykin (2001), a rede neural recorda ao cérebro humano em duas características: 1) o conhecimento é obtido pela rede por meio de seu ambiente, mediante o processo de aprendizagem; 2) A força que uma conexão entre dois neurônios artificiais têm sobre a disseminação de informação entre eles, chamado de pesos sinápticos, sendo cruciais para o funcionamento e a capacidade de aprendizado das redes neurais. Na Figura 2, está ilustrado de forma simplificada, os principais componentes de um neurônio biológico.

Figura 2 – Neurônio biológico



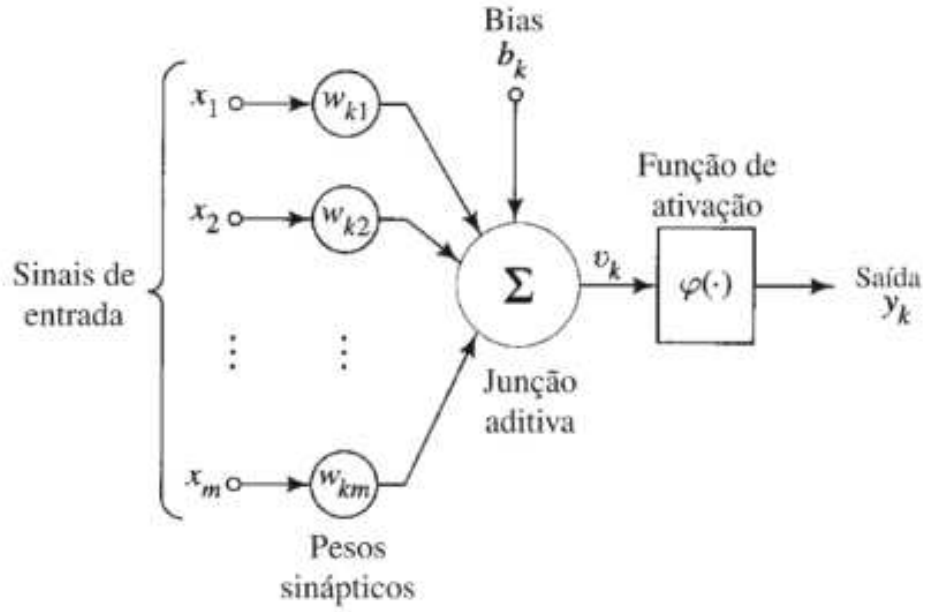
Fonte: (NOVAIS, 2016, p. 45)

O neurônio biológico pode ser dividido em três partes: corpo celular, dendritos e o axônio. Os dendritos recebem as informações dos impulsos nervosos, provindo de outros neurônios, levando-as ao corpo celular, onde é processado as informações criando novos impulsos, esses impulsos são transportados para outros neurônios por meio do axônio que vão de encontro aos dendritos dos neurônios que estão perto. As sinapses também são componentes essenciais para as redes neurais biológicas, podendo exercer uma função excitatória ou inibitória sobre o impulso nervoso, esse controle de conduzir os impulsos faz com que os neurônios se conectam, formando assim as redes neurais biológicas (BRAGA; CARVALHO; LUDERMIR, 2007).

O neurônio artificial teve sua aparição com o modelo (MCP) proposto por Warren McCulloch e Walter Pitts em 1943, onde foi objetivado associar o modelo a um neurônio biológico, no qual possui n entradas, que são representadas pelos dendritos, na qual estas entradas ganham valores x1, x2, x3, . . . ,xn, que refletem os sinais enviados pelos neurônios anteriores. Este modelo possui uma entrada de saída y, retratando o axônio do neurônio biológico. O comportamento das sinapses está no modelo através da conexão dos pesos w1, w2, w3, . . . , wn, com as

entradas, onde esses pesos desenpenham uma sinapse excitatória quando seu valor é positivo, ou sinapse inibitória, quando seu valor é negativo. Abaixo encontra-se a representação gráfica desse modelo, apresentada na Figura 3.

Figura 3 – Modelo de um neurônio artificial



Fonte: (HAYKIN, 2001, p. 36)

De maneira consolidada, as redes neurais possuem características como: generalização, capacidade de aprendizagem e aplicação. A generalização possibilita à rede neural gerar saídas adequadas fundamentadas na sua aprendizagem, isso ocorre quando os dados de entrada são distintos do conjunto de treinamento utilizado na sua aprendizagem. Quando estão sob um processo de treinamento, a capacidade de aprendizagem está em gerar determinadas saídas a partir de determinados padrões de entrada. Na aplicação, a rede se encontra executando a atribuição para qual foi designada, como por exemplo a classificação de padrões, otimização, mapeamento de funções, entre outras (BRAGA; CARVALHO; LUDERMIR, 2007).

O processamento de linguagem natural (PLN) integra um subgrupo na IA que engloba processos de interação da linguagem humana e os computadores, com objetivo de tornar capaz a compreensão e interpretação, gerando uma linguagem natural de maneira útil. Jurafsky e Martin (2008, p. 6) discorrem que "... para muitos, a habilidade de computadores processarem a língua tão bem quanto humanos significará a chegada de máquinas realmente inteligentes".

A linguagem possui muitos aspectos, com isso cada uma delas é abordada de forma distinta no computador. Dias da Silva (2006) discorre sobre cinco níveis de processamento linguístico, sendo eles: fonético-fonológico, morfológico, sintático, semântico e pragmático-discursivo. De maneira sucinta, fonética é o estudo dos sons da fala humana, fundamental para a fonologia, que averigua como os sons funcionam em sistemas linguísticos e também como eles são organizados e utilizados para formar palavras e frases em diferentes idiomas. Tendo

como propósito, verificar como os fonemas se juntam para formar as sílabas gerando palavras e como são acentuadas e pronunciadas. Incorporadas no PLN, ambas, fonética e fonologia, definem um mesmo nível de processamento linguístico.

Liddy (2001) aborda que o computador explora as ondas sonoras e através disso gera uma observação para utilizar posteriormente em alguma aplicabilidade de regras e comparação com modelos da língua. De acordo com Jurafsky e Martin (2008), o reconhecimento da linguagem pode ser por palavras, de maneira individual e também de forma contínua.

No nível morfológico, o processamento de linguagem busca atribuir significado para as palavras, nessa fase é desenvolvido os dicionários das línguas naturais que podem ser legíveis por meio do computador. A tokenização é o método de separar um texto em partes menores, sendo palavras ou frases. Aqui começa uma das muitas aplicações do PLN, permitindo que o sistema processe o texto de maneira ordenada.

No processamento de linguagem natural, a sintaxe "determina o papel de cada uma das palavras de uma sentença e, assim, permite ao sistema convertê-la em estruturas mais facilmente manipuláveis"(COPPIN, 2004, p. 573). Sendo essencial para que os sistemas de inteligência artificial consigam compreender a organização das palavras em uma sentença e, assim, entender o significado que seu usuário estava querendo passar, fazendo tudo isso de forma eficaz, onde após verificar a estrutura gramatical das frases, o sistema de PLN alcançam a interpretação dos dados, pode corrigir e fornecer respostas, isso melhora a interação dos computadores com os seres humanos.

O nível semântico apresenta a compreensão do significado das palavras, frases e textos que estão dentro de um contexto específico. De outro lado, a análise sintática destaca a estrutura gramatical, buscando entender o que as palavras querem dizer e como elas se entrelaçam para formar um significado coerente. No processamento de linguagem natural, a semântica consegue solucionar ambiguidades, pois detecta o significado correto das palavras, já que elas podem ter muitas interpretações dependendo do contexto, classificando características específicas em um texto. Conforme Jurafsky e Martin (2008), o significado de um texto não é baseado somente nas palavras, mas sim no agrupamento e relação que elas apresentam.

O nível pragmático-discursivo apresenta uma interpretação da linguagem enfatizando seu contexto de modo mais amplo, onde relaciona a intenção do usuário e as interações discursivas. Indo além da estrutura e no significado literal das palavras, mostrando como o significado é de certa forma influenciado pelo contexto, pela relação entre as falas anteriores e posteriores, e também pelas condutas culturais e sociais. Tornando-se essencial para uma compreensão mais profunda e precisa da comunicação humana.

Portanto, o processamento de linguagem natural representa um avanço grande na área da inteligência artificial, deixando que máquinas compreendam e interajam com a linguagem humana de maneira ainda mais requintada e natural. Integrando análises fonético-fonológico,

morfológico, sintáticas, semânticas e pragmático-discursivas, o PLN possibilita que os sistemas de IA interpretem não apenas a estrutura das palavras, como também os seus significados e intenções implícitas, a todo momento se adaptando-se ao contexto. Um exemplo disso são os assistentes virtuais, como também análise de sentimentos e tradução automática, tudo de modo muito mais natural, preciso e eficaz.

A junção do aprendizado de máquina, redes neurais e o processamento de linguagem natural marca um avanço muito importante para a inteligência artificial, transformando dados em conhecimentos aplicáveis. O aprendizado de máquina quando se utiliza de redes neurais artificiais, faz com que os sistemas da IA estude volumes grandes de dados e nesse método ocorre o aprimoramento do seu modelo, para melhorar e poder dar continuamente seu desempenho em atividades específicas. Já quando aplicado ao PLN, ele faz com que as máquinas tenham capacidade de entender, produzir e interagir com a linguagem humana de maneira mais eficaz e contextualizada.

3.2 Classificação de documentos

A classificação arquivística é uma atividade que permite organizar os documentos com a finalidade de facilitar a recuperação, o acesso e a gestão da massa documental. No processo de classificação, os documentos são agregados de acordo com um plano de classificação que foi desenvolvido, sendo usualmente representado em uma estrutura, ocorrendo uma hierarquia entre as classes e subclasses. A classificação é um instrumento para organizar documentos desenvolvidos na era contemporânea, apoiando a gestão de documentos das organizações, que se tornam ainda mais complexos com o tempo (GUERCIO, 2002).

A classificação documental é importante para todas as organizações, independentemente de seu tamanho ou setor, pois ela é a chave para a eficiência administrativa e conformidade legal. Mediante análise do organismo produtor de documentos de arquivo, são criadas as classes e divisões direcionadas às funções. Por sequência, as subclasses são incorporadas às tipologias. De acordo com Bellotto (2004, p. 52), a tipologia documental é "a ampliação da diplomática na direção da gênese documental e de sua contextualização nas atribuições, competências, funções e atividades da entidade geradora/acumuladora". Desse modo, "tipologia é o estudo dos tipos documentais (aliando arquivística e diplomática)" Bellotto (2014, p. 349).

Explorando a classificação de documentos, à medida que os documentos ingressam ou saem da organização, ajuda a atualizar o plano de classificação, devendo incluir novos tipos de documentos recebidos ou destinados que não foram previstos no plano, logo, ele será um instrumento imprescindível para a realização da gestão de documentos. Onde o acesso aos documentos se torna facilitado, rápido e eficiente, tanto para os funcionários internos na organização quanto para as partes interessadas externas. Investir em um método eficaz de classificação de documentos se torna essencial para o sucesso da organização, prestando assim um serviço de maior qualidade. Sousa (2007) discorre sobre a classificação como uma função matricial, sendo

o ponto de partida para o desenvolvimento de outras funções como a avaliação e descrição. Resultando a classificação em um elemento essencial para a transparência e o compartilhamento de informações, tornando-se um meio seguro para o exercício da cidadania, preservação da memória e tomada de decisão.

3.3 Classificação funcional

A classificação de documentos de arquivo está presente nas funções em vez da estrutura organizacional, visto que entender as funções é crucial para a compreensão dos documentos. De acordo com Campbell (1941), toda unidade governamental tem suas funções a desempenhar, pois, em caso contrário, não haveria sentido criá-la, visto que todas as vantagens da abordagem organizacional podem ser mantidas e também o classificador pode evitar a compulsão de tentar dividir os documentos em segmentos cronológicos paralelos à vida de cada unidade administrativa envolvida.

Os documentos de arquivo são decorrentes de atividades, funções, processos, entre outros, e esses documentos podem ser usados para representar o ocorrido, mesmo depois de seu encerramento. Os documentos são criados no nível de uma atividade ou de uma transação decorrente dessa atividade e podem ser agregados para formar registros de funções, processos ou outras ocorrências (YEO, 2012). Sem a classificação funcional, ou seja, sem seu vínculo arquivístico, um documento de arquivo não pode ser criado. "Os documentos que não são expressão de uma transação não são documentos de arquivo (records) até que eles sejam colocados dentro de uma relação com outros documentos (records)" (DURANTI, 1997, p. 216). Com isso, a abordagem funcional está no contexto de uso e para qual ele foi gerado, assim sendo, ele será classificado de acordo com a razão de sua criação. A prática de classificar documentos se origina da necessidade de explicitar o "vínculo arquivístico" que existe entre todos os documentos que participam da mesma atividade, desde quando foi gerada.

O National Archives of Australia (2003) entende que a classificação de documentos e informações baseada nas funções e atividades de negócios da instituição se distanciava da classificação tradicional elaborada com base na estrutura organizacional ou por "assuntos". As funções e atividades são mais estáveis do que estruturas organizacionais, pois as funções de uma instituição se mantêm iguais por um grande período de tempo. Se guiando pelo setor público australiano, é argumentado que as mudanças administrativas podem gerar perda ou ganho de funções entre os órgãos, com isso facilitando a identificação de documentos que necessitam seguir as funções.

Algumas observações foram feitas ao praticar o modelo funcional, para elaborar os planos de classificação. Na literatura, Orr (2005) aborda o entendimento que o arquivista pode garantir tendo uma compreensão completa da instituição ou órgão em que esteja atuando, gerando um foco em nível corporativo, a fundamentação da avaliação e a descrição de documentos e consequentemente sua gestão. Com isso, é entendido que a classificação funcional forma o

vínculo arquivístico, onde é possível contextualizar os documentos, dar a opção de ser criar uma estrutura lógica, identificar o que faz parte de um fundo, e ver o que cumpre sua função legal e administrativa, assim como também trazer a recuperação desses documentos. Desse modo, a classificação funcional fornece o controle após a criação do documento, gerando uma gestão integrada de sistemas de documentos de forma híbrida. Contudo, um ponto de vulnerabilidade na classificação funcional são os termos como função, subfunção e atividade no qual mostra uma falha conceitual para a elaboração do plano de classificação.

3.4 Avaliação da usabilidade da classificação funcional

A falta de um consenso sobre o termo função é um problema para a compreensão de metodologias que indiquem os procedimentos para a elaboração de um plano de classificação funcional. Na literatura, há poucas definições sobre função, a falta de uma definição universal é decorrente das diferentes interpretações e variações dentro da arquivologia.

> Qualquer objetivo de alto nível, responsabilidade ou tarefa prescrita como atribuição de uma entidade coletiva pela legislação, política ou mandato. Funções podem ser decompostas em conjuntos de operações coordenadas, tais como subfunções, procedimentos operacionais, atividades, tarefas ou transações (CIA, 2007, p. 13).

As dificuldades na compreensão do conceito de função na área da arquivologia é alto, Eastwood (1994) disserta que estamos longe de entender o que é a função e como ela fundamenta a criação dos documentos. Hurley (1993) declara que não sabe com certeza o que é uma função, porém sabe o que não é, o autor cita um exemplo tratando a função como assunto.

Desse modo, temos que Orr (2005) realizou uma pesquisa sobre as visões dos especialistas e a experiência dos gestores de documentos da Austrália, Canadá e Reino Unido. Foram apresentadas as desvantagens de se aplicar o método funcional de classificação de documentos, foi relatado pelos usuários que o instrumento não era intuitivo, que eles achavam confuso e que por isso ocorria uma diminuição na produtividade com o tempo que era gasto para a compreensão da classificação.

Alguns autores fizeram o levantamento de certas percepções que são apresentadas pelos usuários, Gunnlaugsdottir (2012) aponta que os usuários consideram que os esquemas de classificação funcional são complicados, não é fácil de se utilizar, não é intuitivo e nem tão pouco sabem onde arquivar o documento em um esquema de classificação funcional, visto que é possível classificar o documento em diversas categorias, com isso chegaram a conclusão que o pior método é o funcional, além de seu uso ser demorado eles consideram desnecessário o plano de classificação funcional. Orr (2005) indica também que os usuários não entendem o plano de classificação nem como ele funciona para se ter um bom uso, consequentemente não gostam de usar a classificação funcional, não é esclarecedora nem propriamente eficaz.

Quadro 2 – Percepções dos usuários sobre a classificação funcional

Percepções	**Fontes**
Os usuários consideram que os esquemas de classificação funcional não são intuitivos e nem fáceis de utilizar.	GUNNLAUGSDOTTIR, Johanna. Functional classification scheme for records: a way to chart documented knowledge in organizations. Records Management Journal, v. 22, n. 2, p. 116–129, 2012. JOSEPH, Pauline. EDRMS search behaviour: implications for records management principles and practices, 2010. Tese – University of Western Australia, Perth, 2010.
Os usuários consideram os esquemas de classificação funcional complicados ou demasiadamente complicados.	GUNNLAUGSDOTTIR, Johanna. Functional classification scheme for records: a way to chart documented knowledge in organizations. Records Management Journal, v. 22, n. 2, p. 116–129, 2012. JOSEPH, Pauline. EDRMS search behaviour: implications for records management principles and practices, 2010. Tese – University of Western Australia, Perth, 2010.
Os usuários nem sempre sabem onde arquivar o documento em um esquema de classificação funcional.	GUNNLAUGSDOTTIR, Johanna. Functional classification scheme for records: a way to chart documented knowledge in organizations. Records Management Journal, v. 22, n. 2, p. 116–129, 2012. JOSEPH, Pauline. EDRMS search behaviour: implications for records management principles and practices, 2010. Tese – University of Western Australia, Perth, 2010.
Os usuários acham os esquemas de classificação funcional confusos por ser possível classificar os documentos em muitas categorias.	GUNNLAUGSDOTTIR, Johanna. Functional classification scheme for records: a way to chart documented knowledge in organizations. Records Management Journal, v. 22, n. 2, p. 116–129, 2012. ORR, Stuart Anthony. Functions-based classification of records: is it functional? 2005. Dissertação (Master of Science in Records Management) – Divisão de Informação e Estudos de Comunicação, Northumbria University, Newcastle, 2005. SMYTH, Z. A. Implementing EDRM: has it provided the benefits expected? Records Management Journal, v. 15, n.3, p. 141-149, 2005.
Os usuários acham que o pior método é o funcional.	GUNNLAUGSDOTTIR, Johanna. Functional classification scheme for records: a way to chart documented knowledge in organizations. Records Management Journal, v. 22, n. 2, p. 116–129, 2012.

Os usuários consideram desnecessário o plano de classificação funcional e o uso é demorado.	GUNNLAUGSDOTTIR, Johanna. Functional classification scheme for records: a way to chart documented knowledge in organizations. Records Management Journal, v. 22, n. 2, p. 116–129, 2012.
Os planos de classificação funcional não são intuitivos ou nem amigável para o usuário. Ele fica desconfortável com a abordagem funcional.	JONES, Pauline. The role of virtual folders in developing an electronic document and records management system: Meeting user and records management needs, Records Management Journal, v.18, n.1, p. 53-60, 2008.
Os usuários não compreendem como o plano de classificação funciona e como usar. Não gostam de usar a classificação funcional, porque é difícil de aplicar e é pouco clara.	JOSEPH, Pauline. EDRMS search behaviour: implications for records management principles and practices, 2010. Tese – University of Western Australia, Perth, 2010. ORR, Stuart Anthony. Functions-based classification of records: is it functional? 2005. Dissertação (Master of Science in Records Management) – Divisão de Informação e Estudos de Comunicação, Northumbria University, Newcastle, 2005. PACKALÉN, Saara, HENTTONEN, Pekka, Recordkeeping professional understanding of and justification for functional classification: Finnish public sector organizational context, Archival Science, v. 16, p. 403-419, 2016.
Os usuários consideram a função um conceito estranho e difícil de compreender e não intuitivo.	CALABRIA, Tina. Evaluating Caloundra City Council's EDMS classification. 2004. Disponível em: https://www.steptwo.com.au/files/kmc_caloundracouncil.pdf. Acesso em: 7 mai. 2020. FOSCARINI, Fiorella. Understanding functions: An organizational culture perspective. Records Management Journal, v. 22, p. 20–36, 2012. ORR, Stuart Anthony. Functions-based classification of records: is it functional? 2005. Dissertação (Master of Science in Records Management) – Divisão de Informação e Estudos de Comunicação, Northumbria University, Newcastle, 2005.
Os usuários consideram que os planos de classificação funcional são inflexíveis em relação com as antigas unidades compartilhadas com estrutura pessoal, tornando os planos impopulares.	FOSCARINI, Fiorella. Understanding functions: An organizational culture perspective. Records Management Journal, v. 22, p. 20–36, 2012. MORELLI, Jeff. Hybrid filing schemes: the use of metadata signposts in functional file plans, Records Management Journal, v. 17, n. 1, p. 17-31, 2007.

Para os usuários, a classificação não é intuitiva, é desnecessariamente complexa e mais demorada do que por outros meios.	BAK, Greg. Continuous classification: capturing dynamics relationships among information resources. Archival Science, v. 12, p. 287-318, 2012.
Pelo fato da classificação funcional permitir que um documento seja organizado num único local (fixidez), os usuários também têm capacidade de pesquisa limitada.	ALBERTS, Inge, SCHELLINCK, Jen, EBY, Craig, MARLEAU, Yves. Bridging Functions and Processes for Records Management. Canadian Journal of Information and Library Science, Toronto, v. 34, n. 4, p. 365-390, dez. 2010.

Fonte: adaptado de Ifould e Joseph (2016).

A partir desses levantamentos apresentados pelos autores Ifould e Joseph (2016) se tem uma conclusão sobre a dificuldade que os usuários têm em colocar em prática o plano de classificação funcional. O instrumento deve atender as necessidades e expectativas de seu usuário, pois é utilizado diretamente pelos usuários desde o momento que ele cria ou recebe o documento decorrente de suas atividades no órgão ou instituição.

4 Prova de Conceito: classificação por meio da inteligência artificial

A classificação de documentos sempre foi feita pela humanidade, pois naturalmente o homem sente a necessidade de organizar suas coisas em categorias, a classificação de certo modo está presente em vários momentos da vida dos seres humanos, mesmo que faça sem perceber, é feito uma classificação por um longo tempo na rotina de cada um e isso levou a evolução da sociedade.

Com a expansão do ambiente digital na arquivologia, a ideia inicial seria diminuir as caixas e pastas no espaço físico, entretanto, na classificação de documentos não houve uma otimização que a tecnologia da informação poderia trazer. Na literatura se tem poucos exemplos sobre a classificação automática de documentos de arquivo, é encontrado auto-classification ou automatic classification, essa definição de conceito ainda não se encontra nos dicionários da ciência da informação.

O uso da classificação automática se torna cada vez mais essencial, pois o volume documental crescente nos ambientes digitais requer uma classificação eficiente em relação a grandes quantidades de documentos, visto que, manualmente a classificação torna-se impraticável perante volumes extensos. A classificação manual apresenta muitas inconsistências, seja por falta de padronização ou interpretações subjetivas, desse modo a automação minimiza os erros, obtendo resultados precisos.

A importância da utilização da IA está em proporcionar a automação das tarefas repetitivas. Com isso, é melhorado o sistema de busca e recuperação, facilitando a pesquisa por documentos específicos e informações relevantes em um grande volume de dados, isso promove o acesso às informações de forma ágil. A inteligência artificial pode classificar os documentos tendo como base seu conteúdo e contexto, isso torna a classificação mais abrangente e precisa. Logo, se torna escalável podendo ser adaptada para lidar com diferentes arquivos dentro de cada realidade. Desse modo, a IA está tomando cada vez mais notoriedade na arquivologia, sendo capaz de lidar com grandes volumes de documentos, aumento da eficiência, uma melhor acessibilidade e análise avançada.

Esse trabalho visa, portanto, fazer uma prova de conceito para verificar se é possível utilizar a IA para essa finalidade. E caso a prova de conceito seja bem sucedida, o uso da IA pode ser avaliada e gerar a partir dela uma solução para as dificuldades apresentadas pelos usuários. Os princípios de classificação, no ambiente digital, permanecem os mesmos daqueles utilizados para os documentos em papel, mas os métodos utilizados para aplicá-los podem ser muito diferentes (SHEPHERD; YEO, 2003).

Esse trabalho não tem como objetivo indicar a melhor forma de se treinar um modelo de inteligência artificial, mas sim mostrar que é viável sua utilização para a classificação arquivística. Foi utilizada uma quantidade limitada de documentos, dentro de um contexto específico, em uma situação com mais documentos disponíveis e com a IA sendo implementada dentro de

um sistema informatizado de gestão arquivística de documentos (SIGAD). Essa implementação será feita de forma mais refinada e com muito mais recursos, melhorando ainda mais sua eficiência e podendo lidar com particularidades e situações complicadas que apareçam em um cenário real.

Serão analisados documentos do Senado Federal e o desenvolvimento desta prova de conceito se dá a essas circunstâncias, com isso o uso da inteligência artificial poderá ser utilizada com seus devidos ajustes para cada órgão ou instituição, pois cada uma desenvolve funções e finalidades diferentes. Os documentos selecionados já se encontravam classificados manualmente pelo Senado Federal, eles foram escolhidos com objetivo de se treinar um modelo de aprendizado de máquina capaz de classificar documentos de acordo com seu conteúdo. Esse modelo vai ser utilizado pela inteligência artificial para analisar o conteúdo de um documento e com base nisso classificá-lo na categoria correta. De acordo com o desempenho da inteligência artificial na classificação desses documentos, podemos avaliar se esta é uma alternativa viável a ser utilizada e com isso ter um resultado positivo ou não da prova de conceito.

4.1 Construção da inteligência artificial para a prova de conceito

A construção da IA para a prova de conceito iniciou-se definindo qual seria o funcionamento desejado. O funcionamento escolhido foi de uma inteligência artificial que seria treinada com documentos de um determinado número de categorias diferentes e a partir deste treinamento ela seria capaz de analisar um novo documento e classificá-lo dentro de uma das categorias utilizadas no seu treinamento.

Para a implementação desse funcionamento foi definido um contexto de documentos a serem utilizados: documentos gerados pelo Senado Federal em suas atividades, documentos onde pode-se recuperar o seu conteúdo como texto, visto que, documentos escaneados onde seu conteúdo poderia ser identificado apenas como imagem não foram utilizados, pois a recuperação do seu conteúdo como texto seria mais complexa, necessitando a utilização da técnica de Optical Character Recognition (OCR), ou seja, reconhecimento ótico de caracteres, que identifica caracteres a partir de imagens e isso adicionaria uma complexidade não necessária para o objetivo deste trabalho. Para a seleção de uma unidade de classificação foi levado em consideração as que não possuíam grau de sigilo, a quantidade de categorias escolhidas foi 10 e a quantidade de documentos de cada categoria a serem utilizados para o treinamento da IA foi de 20. Foi definida essa quantidade pois o objetivo desta prova de conceito é mostrar a viabilidade da classificação de documentos utilizando inteligência artificial e não desenvolver a solução mais eficiente, abrangente e otimizada, isso seria um próximo passo a ser feito durante a implementação de uma solução do tipo dentro de um órgão ou instituição, levando em conta suas necessidades e particularidades.

Com o funcionamento desejado definido, foi necessário escolher um algoritmo de clas-

sificação para aprendizado de máquina[1]. Essa escolha levou em conta algoritmos que atendessem dois pontos principais: análise de textos e classificação multiclasse. A análise de texto pois os documentos são em sua maioria textos e a classificação multiclasse, pois esse tipo de classificação consiste em atribuir uma classe, dentre três ou mais, ao objeto que está sendo classificado, o que se encaixa no funcionamento desejado uma vez que temos dez classes. Com isso o algoritmo escolhido foi o Naive Bayes Multinomial, um popular algoritmo de classificação para textos devido a sua eficiência computacional e bom desempenho preditivo (CHEN et al., 2009).

O algoritmo Naive Bayes Multinomial é baseado no teorema de Bayes, um conceito fundamental na teoria das probabilidades, formulado pelo matemático Thomas Bayes no século XVIII. O teorema de Bayes descreve a probabilidade de um evento com base em conhecimento prévio sobre condições relacionadas ao evento. O algoritmo Naive Bayes aplica esse teorema para classificação, assumindo que as características (ou atributos) dos dados são independentes entre si e contribuem de forma isolada para a classificação, por isso o termo naive (ingênuo) em seu nome, já que normalmente há sim uma relação entre as características (BATISTA; BAGATINI; FROZZA, 2018).

Utilizando o algoritmo Naive Bayes Multinomial e um conjunto de 200 documentos do Senado Federal, 20 documentos de cada uma das 10 unidades de classificação escolhidas, o engenheiro de software Jhonatan Alves desenvolveu uma aplicação utilizando a linguagem de programação Python para realizar o treinamento da IA, a avaliação do seu desempenho e a possibilidade de inserir um novo documento para ser classificado. Considere deste ponto em diante, categoria como unidade de classificação.

As 10 categorias estão listadas no Quadro 3 com seus respectivos nomes, descrições e como serão identificadas dentro da IA.

[1] Aprendizado de máquina de máquina é uma forma de alcançar inteligência artificial, ela resolve problemas usando algoritmos e modelos estatísticos para extrair conhecimento de dados e aprender automaticamente (BARKVED, 2022a)

Quadro 3 – Categorias dos documentos utilizados na prova de conceito

Nome	**Identificação**
60.02.01.23 Solicitação de cartão de estacionamento	Categoria 01
53.03.03.16 Relatório da participação em eventos	Categoria 02
54.04.03.07 Requerimento de desarquivamento de documento	Categoria 03
56.04.01.19 Declaração de acumulação de cargo	Categoria 04
56.04.04.13 Portaria de exoneração de servidor	Categoria 05
56.05.09.44 Requerimento de licença para doação de sangue	Categoria 06
56.05.09.51 Requerimento de afastamento para tratar de interesses particulares	Categoria 07
54.04.03.15 Termos de delegação de sigilo	Categoria 08
56.06.02.61 Autorização da chefia do aluno para realização do curso	Categoria 09
58.08.06.22 Termo de recebimento definitivo referente à construção, reforma e manutenção de bem imóvel	Categoria 10

Fonte: elaboração própria.

Para o treinamento da IA a aplicação realiza a leitura dos documentos no formato PDF, que devem estar separados em pastas correspondentes a sua categoria. Após a leitura do documento, é extraído todo o seu texto e é associado à sua categoria. Esse processo é realizado no código retratado na Figura 4, onde os documentos estão organizado em 10 pastas diferentes, correspondentes a cada categoria, e o seu conteúdo é extraído do PDF utilizando a biblioteca Python PyPDF2 que possibilita guardar todo o conteúdo de texto de um arquivo do tipo PDF. Após extrair o conteúdo do PDF, o código armazena esse conteúdo em uma lista e o associa a sua respectiva categoria, que é identificada pela pasta a qual ele pertence.

Figura 4 – Código com a leitura e organização dos documentos

```
# Pasta contendo os PDFs das pastas "tipo0" a "tipo9"
pastas = ['/documentos/tipo0',
          '/documentos/tipo1',
          '/documentos/tipo2',
          '/documentos/tipo3',
          '/documentos/tipo4',
          '/documentos/tipo5',
          '/documentos/tipo6',
          '/documentos/tipo7',
          '/documentos/tipo8',
          '/documentos/tipo9',]

# Lista para armazenar os documentos e categorias
documentos = []
categorias = []

# Itera sobre as pastas de tipo0 a tipo9
for pasta in pastas:

    # Itera sobre os arquivos de cada pasta
    for nome_arquivo in glob(os.path.join(pasta, "*.pdf")):
        caminho_arquivo = os.path.join(pasta, nome_arquivo)

        # Lê o conteúdo do PDF
        with open(caminho_arquivo, 'rb') as arquivo_pdf:
            leitor = PdfReader(arquivo_pdf)
            conteudo = ''
            for pagina in range(len(leitor.pages)):
                conteudo += leitor.pages[pagina].extract_text()

            # Adiciona o conteúdo do documento e a categoria à lista
            numero = pasta.replace('/documentos/tipo', '')
            documentos.append(conteudo)
            categorias.append('Categoria ' + str(numero))
```

Fonte: elaboração própria.

Feita a associação do texto de cada documento com sua respectiva categoria, é realizada uma divisão dos documentos da seguinte forma: 80% dos documentos serão utilizados para o treinamento do modelo Naive Bayes Multinomial e 20% dos documentos serão utilizados para testar o modelo treinado e gerar estatísticas em relação ao seu desempenho. Essa separação é feita pois não podemos testar o modelo utilizando documentos que foram usados no treinamento, já que isso geraria resultados enviesados (BARKVED, 2022b).

Essa separação em dois grupos, treinamento e teste, é realizada utilizando a função *train_test_split*, disponível na biblioteca Scikit-learn, que separa um conjunto de dados categorizados em dois grupos: um para treinamento e outro para teste, da forma que é necessária para o caso em questão e pode ser visto na Figura 5. A biblioteca Scikit-learn, disponível na linguagem de programação Python, começou como um projeto de verão de David Cournapeau no Google Summer of Code, tendo sua primeira versão pública disponibilizada por Fabian Pedregosa, Gael Varoquaux, Alexandre Gramfort e Vincent Michel em 2010 (PEDREGOSA et al., 2011) e é bastante difundida entre cientista de dados e possui algoritmos de aprendizado de máquina e funções que auxiliam na preparação de dados, treinamento e avaliação de modelos

de aprendizado de máquina.

Figura 5 – Código com a separação dos documentos nos grupos de treinamento e teste

```
# Divisão em conjunto de treinamento e teste
documentos_treinamento, documentos_teste, categorias_treinamento, categorias_teste = train_test_split(
    documentos, categorias, test_size=0.2, random_state=0, stratify=categorias
)
```

Fonte: elaboração própria.

Após a separação dos documentos nos grupos de treinamento e teste, o treinamento é realizado utilizando o algoritmo Multinomial Naive Bayes, representado por MultinomialNB() no código da aplicação, e as funções *fit()* e *predict()* da biblioteca Scikit-learn, como demonstrado no código da Figura 6, onde a função *fit()* treina o modelo com base no grupo de documentos de treinamento e a função *predict()* utiliza os documentos do grupo de teste para prever suas categorias, com isso podemos avaliar seu desempenho.

Figura 6 – Código com o treinamento do modelo e a predição utilizando o grupo de teste

```
# Treinamento do modelo Naive Bayes
modelo = MultinomialNB()
modelo.fit(X_treinamento, categorias_treinamento)

# Predição do modelo
categorias_preditas = modelo.predict(X_teste)
```

Fonte: elaboração própria.

A avaliação do desempenho é realizada com base na acurácia. É gerada também uma matriz de confusão, que é uma forma visual de mostrar o número de classificações corretas e incorretas de cada categoria de documentos. Além disso, a aplicação também permite inserir dados manualmente, seja digitando um texto ou inserindo um documento PDF. Esse dado inserido é classificado em uma das 10 categorias, mostrando de forma simplificada como seria o funcionamento da IA em um cenário real onde o usuário do sistema insere um documento e o sistema o classifica automaticamente na categoria correspondente.

Na Figura 7, o código da aplicação escolhe um documento aleatório dentro do grupo de documentos de teste, mostra a qual categoria ele pertence com o texto “Categoria real do documento de teste:” e seu conteúdo com o texto “Conteúdo do documento de teste:”. No final, é utilizada a função *predict()* do Scikit-learn para mostrar qual foi a classificação atribuída ao documento pelo modelo de aprendizado de máquina e o salva na variável *categoria_predita* e a mostra.

Figura 7 – Código com a seleção de um documento aleatório e a predição de sua categoria

```
# Dados de teste
rand = random.randrange(len(documentos_teste))
novo_documento = [documentos_teste[rand]]
print("Categoria real do documento de teste:")
print(categorias_teste[rand])
print("-----------------------------------------------")
print("Conteúdo do documento de teste:")
print(documentos_teste[rand])
print("-----------------------------------------------")

print(type(documentos_teste))

# Vetorização do novo documento
novo_documento_vetorizado = vetorizador.transform(novo_documento)

# Predição do modelo
categoria_predita = modelo.predict(novo_documento_vetorizado)

print('Categoria predita:', categoria_predita)
```

Fonte: elaboração própria.

Por fim, na Figura 8, temos o resultado da execução do código acima, com a categoria real do documento, seu conteúdo e a categoria atribuída pelo modelo de aprendizado de máquina. Com isso, podemos ver que o documento escolhido aleatoriamente pertence à Categoria 7 e foi classificado corretamente pela inteligência artificial, demonstrando o funcionamento da aplicação.

Figura 8 – Execução do código de seleção e classificação de um documento aleatório

```
Categoria real do documento de teste:
Categoria 7
-----------------------------------------------
Conteúdo do documento de teste:

Senado Federal
TERMO DE DELEGAÇÃO DE SIGILO
Eu, [redacted], matrícula [redacted] ocupante do cargo de
DIRETOR DE SECRETARIA, delego a competência para classificar e acessar documentos e
processos no grau de sigilo  PESSOAL – DADOS PESSOAIS , PESSOAL – DADOS DE SAÚDE  e
PESSOAL – JUNTA MÉDICA , de acordo com a Lei nº 12527, de 2011, e o Ato da Comissão Diretora
nº 9, de 2012, no âmbito da(s) unidade(s)  SEIPRE e SEAPOS  para o(s) seguinte(s) colaborador(es):

[redacted] E [redacted], Matrícul a [redacted].

O(s) d elegado(s) declara(m) ter ciência inequívoca da legislação sobre o tratamento de
informação classificada cuja divulgação possa causar risco ou dano à segurança da sociedade ou
do Estado, à intimidade, à vida privada, à honra e à imagem das pessoas e se com promete a
guardar o sigilo necessário, nos termos da Lei nº 12.527, de 2011, e do ATC nº 9, de 2012, e a:
a) tratar as informações classificadas ou os materiais de acesso restrito e preservar o seu
sigilo, de acordo com a legislação vigente;
b) preservar o co nteúdo das informações classificadas, ou dos materiais de acesso restrito,
sem divulgá -lo a terceiros;
c) não praticar quaisquer atos que possam afetar o sigilo ou a integridade das informações
classificadas, ou dos materiais de acesso restrito; e
d) não cop iar ou reproduzir, por qualquer meio ou modo: (i) informações classificadas; (ii)
informações relativas aos materiais de acesso restrito.
Brasília -DF, 30 de janeiro de 2023 .
(Documento assinado eletronicamente pelo delegante e pelo delegado )

CONSULTE EM http://www.senado.gov.br/sigadweb/v.aspx.
ARQUIVO ASSINADO DIGITALMENTE. CÓDIGO DE VERIFICAÇÃO: [redacted].
-----------------------------------------------
<class 'list'>
Categoria predita: ['Categoria 7']
```

Fonte: elaboração própria.

4.2 Resultados da classificação por meio da inteligência artificial

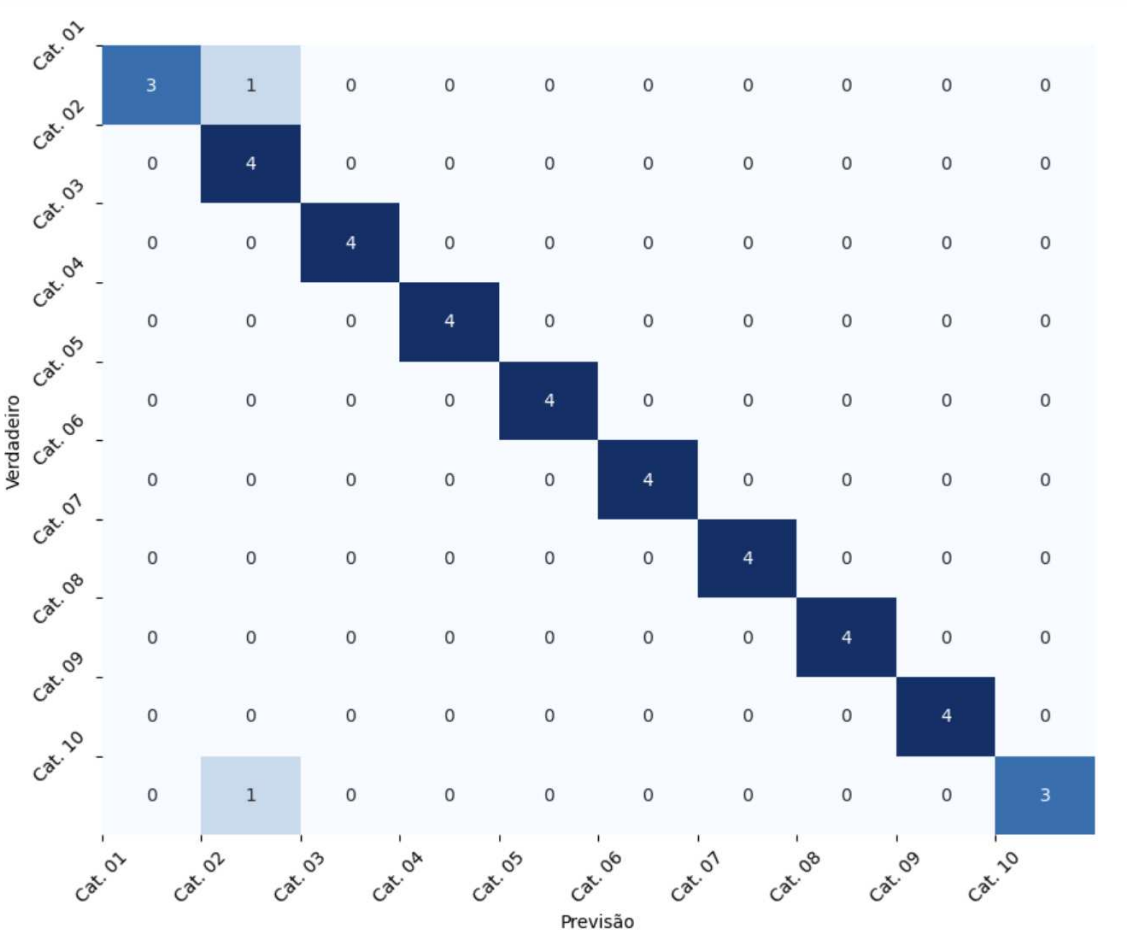
Com a execução da aplicação foram lidos 200 documentos, 20 de cada uma das 10 diferentes categorias. Esses 200 documentos foram separados em dois grupos: documentos para treinamento e documentos para teste. A divisão foi feita em uma proporção de 80% para treinamento e 20% para teste, divisão usualmente utilizada na ciência de dados (BARKVED, 2022b). Com isso, o grupo de documentos para treinamento foi composto de 160 documentos, 16 de cada categoria. Já o grupo de documentos para teste, foi formado por 40 documentos.

Após o treinamento utilizando o grupo de documentos para este fim, o modelo de aprendizado de máquina estava pronto e poderia ser utilizado para classificar documentos dentro de uma das categorias. Assim, o modelo foi utilizado para classificar o grupo de documentos para teste, possibilitando à aplicação avaliar os resultados dessa classificação e gerar estatísticas sobre o desempenho do modelo. O indicador gerado pela aplicação foi a acurácia do modelo, esse indicador mostra um desempenho geral do modelo, sendo calculado pela razão entre o número de classificações corretas e o número total de classificações, conforme a fórmula abaixo:

$$\text{Acuracia} = \frac{\text{total de acertos}}{\text{total de itens}}$$

A acurácia obtida foi de 95%, além do cálculo da acurácia, foi gerada uma tabela chamada matriz de confusão, onde são mostrados os erros e acertos do modelo. Nessa matriz, as linhas correspondem à categoria real do documento classificado e as colunas à categoria predita pelo modelo. O número mostrado na combinação entre uma linha e uma coluna indica quantos documentos pertenciam à categoria da linha e foram classificados na categoria da coluna, se a categoria da linha e da coluna forem iguais temos um acerto, se forem diferentes temos um erro.

Figura 9 – Matriz de confusão



Fonte: elaboração própria.

Na Figura 9, podemos ver que apenas dois documentos foram classificados fora da categoria a qual pertencem, ou seja, classificações incorretas realizadas pela IA. Os demais documentos foram classificados corretamente. Deve-se levar em consideração também que os erros e acertos obtidos são em relação aos documentos que já se encontravam classificados manualmente no Senado Federal.

4.3 Avaliação da prova de conceito

O resultado da prova de conceito foi positivo, pois a inteligência artificial criada teve uma acurácia de 95% na classificação de documentos. Esse número é significativo, pois essa aplicação tem um funcionamento bem simples, sem otimizações e foi utilizada uma quantidade pequena de documentos para seu treinamento. A inteligência artificial criada para a prova de conceito foi capaz de realizar o trabalho de classificação que um ser humano faria e teve um bom desempenho.

Durante a idealização desta prova de conceito, um dos objetivos era avaliar o seu desempenho na classificação de documentos, mas também o custo para atingir um desempenho

satisfatório. Por esse motivo, foi utilizado um modelo simples e com uma quantidade baixa de documentos para seu treinamento, se o desempenho fosse mediano ou ruim, isso levantaria um alerta em relação à aptidão da inteligência artificial na classificação de documentos ou se o custo para atingir bons resultados justifica o benefício. Porém, o bom desempenho mostra que é sim possível utilizar a inteligência artificial na arquivologia para a classificação de documentos, ainda sendo necessário a avaliação em cada contexto específico, porém com uma perspectiva bem positiva dos resultados que podem ser alcançados.

A Figura 10 mostra a classificação de um documento da Caregoria 8, este documento possui 2351 palavras e foi classificado corretamente. A aplicação também mostra o tempo gasto para realizar a leitura e classificação do documento, que foi de aproximadamente 5 milésimos de segundo, um tempo que pode ser considerado insignificante, principalmente comparado ao tempo que uma pessoa levaria para classificar o mesmo documento.

Figura 10 – Quantidade de palavras e tempo de execução

```
Categoria real do documento de teste:
Categoria 8
------------------------------------------
Quantidade de palavras no documento:
2351
------------------------------------------
<class 'list'>
Categoria predita: ['Categoria 8']
Tempo em segundos para conclusão da classificação:
0.00504612922668457
```

Fonte: elaboração própria.

Isso ocorre pois o modelo de inteligência artificial utiliza as probabilidades obtidas no seu treinamento e as aplica sobre o texto a ser classificado, então seu tempo de classificação depende apenas do tamanho do texto e cresce de forma linear. Essa característica é importante por fazer com que a classificação não dependa do tamanho conjunto de dados utilizado no treinamento do modelo, possibilitando assim a utilização de modelos cada vez maiores, se necessário, o que pode dar mais eficiência à classificação, sem impactar no tempo de classificação ao utilizar a inteligência artificial no dia a dia de um órgão ou organização. Esse tempo de classificação dependeria somente do tamanho do documento a ser classificado, porém, como vimos na figura acima, o tempo de classificação para um documento de 2351 palavras foi tão baixo que mesmo para documentos muito maiores isso não seria uma preocupação.

Vale ressaltar que o aumento do tamanho do conjunto de dados utilizados para o treinamento do modelo não tem impacto na classificação de um documento individual, mas impacta sim no tempo de treinamento do modelo. Porém esse treinamento é realizado apenas uma vez para o conjunto de dados, sendo necessário ser realizado novamente apenas em ocasiões onde houver mudanças significativas no perfil dos documentos a serem classificados, forem adicio-

nadas novas categorias ou desejar-se aumentar a eficiência do modelo.

Deve-se ressaltar que para a utilização em um contexto real será necessário o trabalho conjunto dos responsáveis pela implementação da inteligência artificial de forma a integrá-la ao sistema informatizado de gestão arquivística de documentos (SIGAD) utilizado pela organização, além de acompanhamento próximo pelos arquivistas responsáveis, pois os documentos que serão utilizados no treinamento do modelo de inteligência artificial e suas respectivas categorias são parte fundamental para um bom desempenho da solução de classificação automatizada.

Com isso, podemos concluir que a prova de conceito deu uma perspectiva bastante positiva para a utilização da inteligência artificial na classificação de documentos e que sua integração no fluxo de trabalho de gestão documental é sim viável, podendo ser estudada a sua implementação dentro de cada organização.

5 Considerações Finais

Foi mostrada a viabilidade da realização de classificação arquivística por meio de IA, a utilização desse recurso é valioso para enfrentar os desafios de gerenciar as informações em uma era digital em constante crescimento. Sabendo-se dessa viabilidade é possível o desenvolvimento de uma solução dentro do contexto de cada órgão ou instituição que deseje aprimorar a classificação arquivística e evitar os problemas da classificação funcional, já que a solução utilizando inteligência artificial tem potencial de classificar documentos com uma alta acurácia, como demonstrado na prova de conceito.

Desse modo, temos que o uso da inteligência artificial para a classificação de documentos arquivísticos apresenta inúmeras vantagens e benefícios. A IA permite uma análise mais rápida e eficiente dos documentos, melhorando a classificação e otimizando o tempo que os usuários gastam neste processo. É feito o uso de algoritmos e modelos de aprendizado de máquina para que seja feita uma análise e compreensão dos documentos, isso faz com que sejam identificados padrões e recursos que permitam cada vez mais sua classificação correta. Entre os benefícios está apresentado o poder de lidar com grandes volumes de massa documental, essa capacidade é especialmente vantajosa para as organizações que atuam com uma quantidade massiva de informações, como arquivos históricos, instituições governamentais além de empresas com altos fluxos documentais.

Destaca-se a precisão e a velocidade disponibilizada pela inteligência artificial na classificação de documentos, visto que a classificação manual está sujeita a erros humanos, inconsistências e subjetividades. Este ponto garante uma organização confiável e precisa dos documentos arquivísticos, contribuindo para a preservação e acesso efetivo aos registros ao longo do tempo. No entanto, é fundamental ressaltar que o arquivista desempenha um papel essencial na definição dos critérios e parâmetros que serão utilizados pela inteligência artificial, além de solicitar revisões e ajustes necessários para garantir a exatidão da classificação de documentos. Embora a IA seja uma ferramenta poderosa, ela requer um conhecimento especializado do arquivista para garantir que seja utilizada de forma adequada e eficaz.

O papel do arquivista está em garantir que dentro da classificação automática seja respeitado algumas noções importantes como:

- princípio da proveniência, garantindo a relação contextual entre os documentos, sendo assim preservando a integridade dos conjuntos documentais;
- princípio de respeito a ordem original, onde os documentos devem ser organizados conforme foram produzidos e recebidos, sem nenhuma alteração em sua sequência original;
- contexto histórico, a classificação feita por meio da IA deve levar em consideração o período que foram criados os documentos, isso auxilia para uma organização que reflita

a interpretação e entendimento correto dos documentos dentro do contexto que foram produzidos;

- metadados, deve ser atribuído metadados relevantes aos documentos, como por exemplo os dados de criação, autor, dentre outras, isso facilita sua descrição, identificação e recuperação;
- acesso, permitir um acesso eficiente aos documentos e para que isso ocorra, tem que se obter uma estrutura para essa classificação de forma clara e consistente de modo que facilite ao usuário sua busca de forma rápida
- e, por fim, é necessário uma revisão e verificação dos resultados, pois embora a classificação seja realizada por algoritmos, é importante garantir a qualidade da classificação e a correção de falhas.

O contexto e o conteúdo dos documentos desempenham um papel fundamental na classificação automática, pois vão de encontro à capacidade da inteligência artificial em entender e classificar os documentos de forma precisa. No contexto, a semântica traz para a IA o entendimento do significado das palavras em um determinado contexto, permitindo classificar os documentos tendo como base o seu real significado, não apenas utilizando palavras-chave separadas, isso tudo coopera para que a inteligência artificial compreenda a estrutura e as ligações entre as informações.

Essa abordagem é uma ferramenta valiosa para a classificação de documentos. Dentro desses novos comportamentos informacionais os usuários que criam documentos dentro do Sistema Informatizado de Gestão Arquivística de Documentos, buscam uma melhoria para com a classificação funcional, onde esperam que suas demandas e expectativas sejam contempladas dentro desse sistema. Visto que a literatura aponta que a classificação de documentos de arquivo feita pela função apresenta dificuldades para ser um instrumento adequado e prático.

Referências

BARKVED, K. *The difference between AI, machine learning, and deep learning*. 2022. Accessed: 2023-06-13. Disponível em: <https://www.obviously.ai/post/machine-learning-vs-artificial-intelligence-vs-deep-learning-whats-the-difference>. Citado na página 31.

BARKVED, K. *The difference between training data vs. test data in machine learning*. 2022. Accessed: 2023-06-13. Disponível em: <https://www.obviously.ai/post/the-difference-between-training-data-vs-test-data-in-machine-learning>. Citado 2 vezes nas páginas 33 e 36.

BATISTA, R. de A.; BAGATINI, D. D. S.; FROZZA, R. Classificação automática de códigos ncm utilizando o algoritmo naïve bayes. *iSys - Brazilian Journal of Information Systems*, v. 11, n. 2, p. 4–29, 2018. Citado na página 31.

BELLOTTO, H. L. *Arquivos permanentes. Tratamento documental*. 2. ed. rev. e aum. ed. Rio de Janeiro: FGV, 2004. Citado na página 23.

BELLOTTO, H. L. *Arquivos permanentes: tratamento documental*. 4. ed.. ed. Rio de Janeiro: Editora Fundação Getúlio Vargas, 2006. Citado na página 16.

BELLOTTO, H. L. *Arquivo: estudos e reflexões*. Belo Horizonte: UFMG, 2014. Citado na página 23.

BRAGA, A. de P.; CARVALHO, A. C. P. L. F.; LUDERMIR, T. B. *Redes neurais artificiais: teoria e aplicações*. 2ª. ed. [S.l.]: Livros Técnicos e Científicos, 2007. Citado 2 vezes nas páginas 20 e 21.

CAMPBELL, E. G. Functional classification of archival material. *The Library Quarterly*, University of Chicago Press, v. 11, n. 4, p. 431–441, 1941. Citado na página 24.

CHEN, J. et al. Feature selection for text classification with naïve bayes. *Expert Systems with Applications*, v. 36, n. 3, p. 5432–5435, 2009. Citado na página 31.

CONSELHO INTERNACIONAL DE ARQUIVOS. *ISDF – Norma internacional para descrição de funções*. Rio de Janeiro: Arquivo Nacional, 2007. Citado na página 25.

COPPIN, B. *Artificial Intelligence Illuminated*. [S.l.]: Jones and Bartlett Publishers Inc., 2004. Citado na página 22.

DIAS DA SILVA, B. C. O estudo lingüístico-computacional da linguagem. *Letras de Hoje*, EDIPUCRS, Porto Alegre, Brasil, v. 41, n. 144, p. 103–138, 2006. Citado na página 21.

DURANTI, L. The archival bond. *Archives and Museum Informatics*, v. 11, p. 213–218, 09 1997. Citado na página 24.

E-ARQ BRASIL. *Modelo de Requisitos para Sistemas Informatizados de Gestão Arquivística de Documentos*. Rio de Janeiro: Arquivo Nacional, 2011. Citado na página 17.

EASTWOOD, T. What is archival theory and why is it important? *Archivaria*, v. 37, p. 122–130, Jan. 1994. Disponível em: <https://archivaria.ca/index.php/archivaria/article/view/11991>. Acesso em: 26 jan. 2023. Citado na página 25.

FACELI, K. et al. *Inteligência artificial: uma abordagem de aprendizado de máquina*. Rio de Janeiro: LTC, 2011. Citado na página 19.

GUERCIO, M. Records classification and content management: old functions and new requirements in the legislations and standards for electronic recordkeeping systems. In: *Proceedings of the DLM-Forum 2002. Access and preservation of electronic information: best practices and solutions*. Barcelona: [s.n.], 2002. Citado na página 23.

GUNNLAUGSDOTTIR, J. Functional classification scheme for records: Fcs: A way to chart documented knowledge in organizations. *Records Management Journal*, v. 22, p. 116–129, 07 2012. Citado 2 vezes nas páginas 13 e 25.

HAYKIN, S. *Redes neurais princípios e práticas*. 2ª. ed. São Paulo: BOOKMAN, 2001. Citado 2 vezes nas páginas 20 e 21.

HEREDIA HERERRA, A. *Archivistica general. Teoria y practica*. Sevilha: Diputación de Sevilla, 1991. Citado na página 12.

HORTA, D.; CAMPELLO, R. J. G. B. Automatic aspect discrimination in data clustering. *Pattern Recognition*, v. 45, n. 12, p. 4370–4388, 2012. Citado na página 19.

HURLEY, C. What, if anything, is a function? *Archives & Manuscripts*, v. 21, n. 2, 1993. Citado na página 25.

IFOULD, P.; JOSEPH, P. User difficulties working with a business classification scheme: a case study. *Records Management Journal*, v. 26, n. 1, 2016. Citado na página 28.

INNARELLI, H. C. Preservação digital: a gestão e a preservação do conhecimento explícito digital em instituições arquivísticas. *InCID: Revista de Ciência da Informação e Documentação*, Ribeirão Preto, v. 3, n. 2, p. 48–63, 2012. Acesso em: 11 de agosto de 2024. Disponível em: <https://www.revistas.usp.br/incid/article/view/48653>. Citado na página 18.

INTERNATIONAL RECORDS MANAGEMENT TRUST. *Glossary of terms*. London: IRMT, 2009. Citado na página 16.

JURAFSKY, D.; MARTIN, J. H. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition*. 2ª. ed. [S.l.]: Prentice Hall, 2008. Citado 2 vezes nas páginas 21 e 22.

KURZWEIL, R. *The Age of Spiritual Machines*. Massachusetts: The MIT Press, 1990. Citado na página 18.

LIDDY, E. D. Natural language processing. In: *Encyclopedia of Library and Information Science*. 2nd. ed. NY: Marcel Decker, Inc., 2001. Citado na página 22.

MARTÍN-POZUELO CAMPILLOS, M. P. *La construccion teórica en archivistica: el principio de procedencia*. Madrid: Boletin Oficial del Estado, 1996. Citado na página 10.

MCCULLOCH, W.; PITTS, W. A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*, v. 5, p. 115–133, 1943. Citado na página 20.

MOKHTAR, U.; YUSOF, Z. Function-based classification: Model development and validation. *Open Journal of Social Sciences*, v. 3, n. 3, 2015. Citado na página 13.

MONARD, M. C.; BARANAUSKAS, J. A. *Aplicações de Inteligência Artificial: Uma Visão Geral*. São Carlos: Instituto de Ciências Matemáticas e de Computação de São Carlos, 2000. Citado na página 19.

National Archives of Australia. *Overview of Classification Tools for Records Management*. Canberra: National Archives of Australia, 2003. Disponível em: <https://www.naa.gov.au/sites/default/files/2019-10/classifcation-tools.pdf>. Acesso em: 26 jan. 2023. Citado na página 24.

NOVAIS, J. P. Aplicação dos algoritmos sift e surf na classificação de subimagens por discriminação de textura. Marília, 2016. Citado na página 20.

ORR, S. *Functions-based classification of records: is it functional?* Dissertação (Master of Science in Records Management) — School of Informatics, Division of Information & Communication Studies, Northumbria University, Newcastle, 2005. Disponível em: <https://static1.squarespace.com/static/5a1c710fbce17620f861bf47/t/5b6cb3d540ec9ad5dd96d738/1533850585919/Functions-based_classification_of_records_is_it_fu.pdf>. Acesso em: 26 jan. 2023. Citado 2 vezes nas páginas 24 e 25.

PEDREGOSA, F. et al. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, v. 12, p. 2825–2830, 2011. Citado na página 33.

PHAN, T. V.; NAKAGAWA, M. Combination of global and local contexts for text/non-text classification in heterogeneous online handwritten documents. *Pattern Recognition*, v. 51, p. 112–124, Mar 2016. Citado na página 19.

RIOS, E. R.; CORDEIRO, R. I. de N. Plano de classificação de documentos arquivísticos e a teoria da classificação: uma interlocução entre domínios de conhecimentos. perspectivas em ciência da informação. v. 15, n. 2, 2010. Disponível em: <http://hdl.handle.net/20.500.11959/brapci/35509>. Acesso em: 12 jan. 2023. Citado na página 12.

SANTOS, C. N. dos. *Aprendizado de máquina na identificação de sintagmas nominais: o caso do português brasileiro*. Dissertação de Mestrado em Sistemas e Computação, 2005. Citado na página 16.

SCHELLENBERG, T. R. *Documentos Públicos e Privados: Arranjo e Descrição*. 9. ed. Rio de Janeiro: FGV, 1980. Citado na página 11.

SCHELLENBERG, T. R. *Arquivos Modernos: Princípios e Técnicas*. 4. ed. Rio de Janeiro: Fundação Getulio Vargas, 2004. Citado 3 vezes nas páginas 10, 11 e 12.

SHEPHERD, E.; YEO, G. *Managing records: a handbook of principles and practice*. London: Facet, 2003. Citado na página 29.

SOUSA, R. T. B. Os princípios arquivísticos e o conceito de classificação. In: RODRIGUES, G. M.; LOPES, I. L. (Org.). *Organização e representação do conhecimento na perspectiva da Ciência da Informação*. Brasília-DF: Thesaurus, 2003. v. 2, p. 240–269. Citado na página 16.

SOUSA, R. T. B. *Classificação em arquivística: trajetória e apropriação de um conceito*. Tese (Doutorado em História Social) — Universidade de São Paulo, São Paulo, 2005. Citado na página 16.

SOUSA, R. T. B. de. A classificação como função matricial do que-fazer arquivístico. In: SANTOS, V. B. dos; INNARELLI, H. C. (Ed.). *Arquivística temas contemporâneos: classificação, preservação digital, gestão do conhecimento*. Brasília: SENAC, 2007. Citado na página 23.

SPöRL, C.; CASTRO, E. G.; LUCHIARI, A. Aplicação de redes neurais artificiais na construção de modelos de fragilidade ambiental. *Revista do Departamento de Geografia*, v. 21, n. 1, p. 113–135, 2011. Citado na página 20.

YEO, G. Bringing things together: Aggregate records in a digital age. *Archivaria*, v. 74, p. 43–91, Nov. 2012. Disponível em: <https://archivaria.ca/index.php/archivaria/article/view/13407>. Acesso em: 24 jan. 2023. Citado na página 24.