UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Uso de recuperação de informações para auxílio à elaboração de peças de processos do TCU

Marcos Tibúrcio dos Santos Tabosa

Monografia - MBA em Inteligência Artificial e Big Data

ICMC USP
SÃO CARLOS

Marcos Tibúrcio dos Santos Tabosa

Uso de recuperação de informações para auxílio à elaboração de peças de processos do TCU

Monografia apresentada ao Departamento de Ciências de Computação do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo - ICMC/USP, como parte dos requisitos para obtenção do título de Especialista em Inteligência Artificial e Big Data.

Área de concentração: Inteligência Artificial

Orientador: Prof. Dr. Roney Lira de Sales Santos

Versão original

São Carlos

2024

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi, ICMC/USP, com os dados fornecidos pelo(a) autor(a)

Tabosa, Marcos
S856m Uso de recuperação de informações para auxílio à elaboração de peças de processos do TCU / Marcos Tibúrcio dos Santos Tabosa ; orientador Roney Lira de Sales Santos. – São Carlos, 2024.
82 p. : il. (algumas color.) ; 30 cm.

Monografia (MBA em Inteligência Artificial e Big Data) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2024.

1. Inteligência Artificial. 2. Recuperação de informação. 3. Embedings. 4. BERT. 5. Sentence-BERT. 6. Similaridade por Cosseno. 7. Tribunal de Contas da União. I. Santos, Roney L. de S., orient. II. Título.

Marcos Tibúrcio dos Santos Tabosa

Monograph presented to the Departamento de Ciências de Computação do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo - ICMC/USP, as part of the requirements for obtaining the title of Specialist in Artificial Intelligence and Big Data.

Concentration area: Artificial Intelligence

Advisor: Prof. Dr. Roney Lira de Sales Santos

Original version

São Carlos

2024

AGRADECIMENTOS

Agradeço à pequena Elis cuja chegada brindou esse período de estudos e, especialmente, agradeço à minha esposa, Lilian, pela dedicação e carinho perante as novas emoções e responsabilidades.

RESUMO

TABOSA, Marcos T. S. . 2024. 82 p. Monografia (MBA em Inteligência Artificial e Big Data) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2024.

O uso de inteligência artificial tem se demonstrado de grande valia em diversos processos de trabalho. No âmbito do Tribunal de Contas da União (TCU), apesar das recentes iniciativas voltadas ao uso de novas tecnologias, tarefas atinentes à análise de questões de competência do referido órgão ainda carecem de uma ferramenta que possibilite recuperar, de forma eficiente, elementos textuais de deliberações expedidas pelo Tribunal (tais como argumentos, critérios e encaminhamentos) com potencial de servir ao corpo técnico para análise de questões já enfrentadas anteriormente. Neste trabalho, foram utilizadas técnicas de processamento de linguagem natural para construção de uma solução que, baseada em similaridade semântica, identifica e resgata trechos de deliberações do TCU relacionados a determinado assunto objeto de interesse. Em que pese a ausência de critérios estritamente objetivos para avaliar a solução implementada, com base nas simulações realizadas, foram obtidos indicativos de que a ferramenta desenvolvida tem o potencial de contribuir com o aumento da produtividade, uniformidade e coesão do exame de processos finalísticos do TCU.

Palavras-chave: Inteligência Artificial. Recuperação de informação. Embedings. BERT. Sentence-BERT. Similaridade por Cosseno. Tribunal de Contas da União.

ABSTRACT

TABOSA, Marcos T. S. . 2024. 82 p. Monograph (MBA in Artificial Intelligence and Big Data) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2024.

The use of artificial intelligence has proven to be of great value in various work processes. Within the scope of the Tribunal de Contas da União (TCU), despite recent initiatives aimed at the use of new technologies, tasks related to the analysis of issues within the jurisdiction of that body still lack a tool that makes it possible to efficiently retrieve textual elements from deliberations issued by the Court (such as arguments, criteria and referrals) with the potential to serve the technical staff to analyze issues previously faced. In this work, natural language processing techniques were used to build a solution that, based on semantic similarity, identifies and retrieves excerpts from TCU deliberations related to a given subject of interest. Despite the absence of strictly objective criteria to evaluate the implemented solution, based on the simulations carried out, indications were obtained that the tool developed has the potential to contribute to increasing productivity, uniformity and cohesion in the examination of TCU's finalistics processes.

Keywords: Artificial Intelligence. Information Retrieval. Embedings. BERT. Sentence-BERT. Cosine similarity. Tribunal de Contas da União.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

AudPortoFerrovia	Unidade de Auditoria Especializada em Infraestrutura Portuária e Ferroviária
AudRodoviaAviação	Unidade de Auditoria Especializada em Infraestrutura Rodoviária e Aviação Civil
BERT	Bidirectional Encoder Representations from Transformers
DSM	Distributional Semantic Models
FFN	FeedForward Network
GPT	Generative Pre-Trained Transformer
IA	Inteligência Artificial
IR	Information Retrieval
LLM	Large Language Models
MSD	Modelos Semânticos Distribucionais
NER	Named Entity Recognition
NLG	Natural Language Generation
NLP	Natural Language Processing Natural
NLTK	Natural LanguageToolkit
NLU	Language Understanding
PLN	Processamento de Linguagem Natural
RAG	Retrieval-Augmented Generation
RNN	Recurrent Neural Network
RSLP	Removedor de Sufixos da Língua Portuguesa
SeinfraPortoFerrovia	Secretaria de Fiscalização de Infraestrutura Portuária e Ferroviária
SeinfraRodovia-Aviação	Secretaria de Fiscalização de Infraestrutura Rodoviária e de Aviação Civil

SeinfraUrb	Unidade de Auditoria Especializada em Infraestrutura Urbana e Hídrica
SeinfraUrbana	Secretaria de Fiscalização de Infraestrutura Urbana
TCU	Tribunal de Contas da União
TF-IDF	Term Frequency – Inverse Document Frequency

SUMÁRIO

1 INTRODUÇÃO

1.1 Contextualização e Motivação

O poder público, em praticamente todas as suas áreas de atuação, é célebre pela produção de grandes volumes de informação em forma de texto. Essa característica da burocracia estatal decorre especialmente da necessidade de atender aos princípios da publicidade, transparência e formalismo[1] a que estão submetidos os atos públicos[2].

Documentos textuais constituem instrumentos de apoio às funções governamentais, contribuindo para a preservação de sua memória e democratização do acesso às informações públicas. Nesse cenário, adquire grande relevância a gestão de documentos e informações com o fim de dotar o Estado de eficiência e transparência perante o cidadão (Santos e Torres, 2019).

Neste contexto, tem-se que as possibilidades de melhoria dos processos, do tratamento de informações e da viabilização da proteção e testemunho da atuação pública, mediante controle e monitoramento de documentos e arquivos, transpuseram suas finalidades a um patamar de considerável importância (Cortes 1996).

No entanto, gerir o grande volume de documentos e informações que orientam a atuação da máquina pública constitui um relevante desafio no qual ganham destaque os recursos advindos da tecnologia da informação – incluindo o uso de Inteligência Artificial (IA) e Bigdata.

Apesar de geralmente constarem em documentos com tipologias predefinidas, pode-se afirmar, seguramente, que a maior parte da informação textual produzida pela Administração Pública não se encontra estruturada de modo a permitir o fácil acesso para o público em geral ou para o consumo do próprio órgão público que a produziu.

Essa dificuldade de o próprio órgão acessar a informação que produziu, prejudica sua atuação nas dimensões (i) do planejamento (dificultando o direcionamento racional

---

[1] Esses princípios, consagrados pela doutrina do direito administrativo, são regulados ou orientam diversos dispositivos legais expedidos pela União, Estados e Municípios. Em âmbito federal, pode-se citar: (i) publicidade: art. 37 da Constituição Federal e art. 2º, parágrafo único, inciso V da Lei 9.784/1999; (ii) transparência: Lei 12.527/2011 (Lei de Acesso à informação); e (iii) formalismo: art. 60 da Lei 8.666/1993, Lei 9.784/1999 e diversas legislações específicas.

[2] Ato administrativo é toda manifestação unilateral de vontade da Administração Pública, que, agindo nessa qualidade, tenha por fim imediato adquirir, resguardar, transferir, modificar, extinguir e declarar direitos, ou impor obrigações aos administrados ou a si própria. (...) O exame do ato administrativo revela nitidamente a existência de cinco requisitos necessários à formação, a saber: competência, finalidade, forma, motivo e objeto. Tais componentes, pode-se dizer, constituem a infraestrutura do ato administrativo, seja ele vinculado ou discricionário, simples ou complexo, de império ou de gestão (Meirelles *et al.* 1991)

de recursos a distintas demandas), (ii) da produtividade (exigindo revisões e retrabalho desnecessários) e (iii) da homegeneidade da resposta (ensejando ações variadas para situações similares). Dessas observações parte a motivação para elaboração da presente pesquisa.

Da vasta estrutura administrativa brasileira, interessa especificamente a este trabalho a produção documental do Tribunal de Contas da União (TCU), órgão de controle externo do governo federal que auxilia o Congresso Nacional nas missões de acompanhar a execução orçamentária e financeira do país e contribuir com o aperfeiçoamento da Administração Pública.

Anota-se que, além das competências constitucionais e privativas do TCU que estão estabelecidas na Constituição Federal (Brasil 1988), outras leis específicas conferem outras atribuições ao referido Tribunal, dentre as quais estão a Lei de Responsabilidade Fiscal; Lei Complementar 101/2000 (Brasil 2000); as leis que regulam de licitações e contratações públicas, a exemplo da Lei 8.666/1993 (Brasil 1993), 14.133/2021 (Brasil 2021) e 13.303/2016 (Brasil 2016); e, anualmente, a Lei de Diretrizes Orçamentárias[3].

Internamente, assim como se verifica em toda Administração Pública, o TCU é estruturado em unidades organizacionais às quais são atribuídos assuntos, competências ou temáticas específicas. No âmbito do TCU, a atividade de controle externo exercida pelo órgão é realizada pelas denominadas Unidades de Auditoria Especializadas, cujas competências são definidas com base na destinação dos recursos fiscalizados, a exemplo de investimentos das áreas de saúde, educação, cultura e obras de infraestrutura.

É natural que as análises de processos sob responsabilidade das citadas unidades organizacionais se debrucem sobre determinados aspectos técnicos, legais, jurídicos e doutrinários, bastante recorrentes – consequência inerente da especialização por temáticas. Assim, frequentemente, no âmbito destas unidades organizacionais são elaboradas peças processuais em que argumentos, fundamentos e critérios de análise são bastante similares aos já utilizados anteriormente em processo distinto ou mesmo em análises precedentes do mesmo processo.

Por outro lado, ainda que a divisão de trabalho seja pautada por temáticas específicas, subsiste a necessidade de lidar na prática cotidiana com uma ampla gama de situações, assuntos e legislações. Ademais, no âmbito das unidades finalísticas, as análises requeridas nem sempre são delegadas aos servidores que já a trataram de questões similares anteriormente.

[3] No momento deste trabalho, a Lei de Diretrizes Orçamentárias (LDO) que rege o exercício vigente é a Lei 4.791, de 29 de dezembro de 2023, disponível em https://www.gov.br/planejamento/pt-br/assuntos/orcamento/orcamentos-anuais/2024/ldo/lei-de-diretrizes-orcamentarias-ldo

Deve-se anotar que atualmente todos os processos do TCU são mantidos em sistema informatizado (denominado eTCU), que realiza a gestão de todas as peças e eventos processuais relacionados (como trâmites, comunicações, pedidos de vista etc.). No citado sistema também são geridas as peças elaboradas pelo próprio corpo técnico do Tribunal nas quais são formalizadas as análises relativas às atribuições finalísticas do órgão.

Ocorre que o grande volume de documentação associado a cada processo representa real obstáculo à localização de elementos textuais (assim entendidos como tópicos, frases ou parágrafos) que poderiam servir de auxílio à análise de questões recorrentes – situação ainda mais evidente em processos que já superaram algumas das suas diversas etapas e, invariavelmente, acumulam centenas de documentos em seu escopo.

Encontram-se disponíveis, ao público interno e externo, ferramentas dedicadas à busca textual, baseada em palavras-chave, na base de jurisprudência do TCU (denominada "Pesquisa Integrada"[4]). No entanto, não existe uma solução específica voltada a recuperar, de forma prática, análises técnicas anteriores relacionadas à essência de determinada matéria enfrentada no âmbito de um novo processo.

Desta forma, o problema ao qual se dedicou o presente trabalho diz respeito à inexistência de soluções eficientes de busca e compartilhamento de informações sobre análises de assuntos tratados em momento anterior no âmbito do TCU – situação que está associada (i) à perda de produtividade, em razão do retrabalho na construção de análises sobre temas recorrentes, e (ii) maior heterogeneidade no exame e nos encaminhamentos conferidos a questões idênticas, avaliadas em distintos processos.

1.2 Hipóteses

No panorama apresentado, o presente trabalho assumiu como hipótese que as técnicas de Processamento de Linguagem Natural (PLN) podem contribuir para facilitar o acesso, de forma eficiente, a trechos de documentos técnicos potencialmente úteis para o exame de questões recorrentemente enfrentadas no âmbito dos processos de controle externo do TCU.

Assim, a solução desenvolvida neste trabalho baseou-se em ferramenta de captura do conteúdo semântico de parágrafos (ou mesmo pequenas sequências de palavras) de textos fornecidos como entrada, para a identificação de partes de outras deliberações já expedidas pelo Tribunal que abordaram assuntos similares ou correlatos.

[4] O TCU disponibiliza ao público externo quatro bases de dados para pesquisa de sua jurisprudência: Acórdãos, Jurisprudência Selecionada, Publicações, Súmulas, com links disponíveis em https://portal.tcu.gov.br/jurisprudencia/, também é possível encontrar similares informações em ferramenta mais acessível ao usuário em https://pesquisa.apps.tcu.gov.br/pesquisa/integrada

1.3 Objetivos

Em essência, o presente trabalho se voltou ao desenvolvimento de uma solução de auxílio à escrita de análises de processos, que propicia o resgate de argumentos, critérios e encaminhamentos de produção textual anterior relacionada a determinado assunto submetido ao exame de uma unidade técnica do TCU.

Nessa esteira, o objetivo geral do presente trabalho foi desenvolver um algoritmo com o intuito de auxiliar o exame de temas comuns aos processos de controle externo do TCU; resultado obtido por meio de uma solução baseada em processamento de linguagem natural que, a partir de sentenças sobre determinada questão técnica ou jurídica, identifique (e sugira) trechos de análises anteriores correlacionadas ao assunto tratado.

O atingimento dos propósitos apresentados está vinculado à resolução dos seguintes objetivos específicos:

- definições acerca da quantidade e características dos textos que compuseram os repositórios (*datasets*) de deliberações do TCU, dos quais são resgatados os textos que têm potencial de serem "reutilizados" ou "readaptados" nas novas análises;

- pesquisa e escolha de alternativas tecnológicas adequadas para implementação, armazenamento dos *datasets* de documentos textuais, a exemplo da estrutura dos dados manipulados;

- testes e decisões acerca das opções para pré-processamento dos textos dos citados repositórios de textos técnicos, no que se incluiu o tratamento de *stopwords*, as especificações de *tokenização*, bem como a utilização de técnicas de *Stemming*;

- implementação de modelo de identificação de similaridade semântica entre elementos do texto de entrada (frases ou sequências de palavras sobre determinada questão) e os textos do repositório de documentos produzidos no âmbito do TCU;

- implementação de algoritmo para apresentação de forma rápida e assertiva das sentenças do repositório identificadas pela solução, que também possibilite ao usuário acesso a partes dos documentos contíguas às referidas sentenças sugeridas – com o fim de fornecer informações sobre o contexto original em que as sentenças foram utilizadas.

1.4 Organização do trabalho

Ao todo, o conteúdo do presente trabalho é dividido em cinco capítulos. Sucedendo este primeiro capítulo introdutório, o Capítulo 2 tem por propósito assentar a fundamentação teórica da solução utilizada nesta pesquisa. Após a apresentação dos principais conceitos relacionados a dados, informação e conhecimento, esse segundo capítulo traz um panorama das principais características e dos avanços das áreas da Inteligência Artificial (IA) e do Processamento de Linguagem Natural (PLN), com enfoque especial nos

processos envolvidos na presente investigação.

O Capítulo 3 trata da metodologia adotada para elaboração da solução implementada, no se inclui o detalhamento de procedimentos utilizados para construção, tratamento e transformações das bases de dados textuais constituídas a partir de repositório de deliberações do TCU. Nesse capítulo são delineadas as especificações do modelo utilizado para tarefa de recuperação de informações – que constitui o instrumento de operacionalização dos aspectos teóricos relacionados à presente pesquisa.

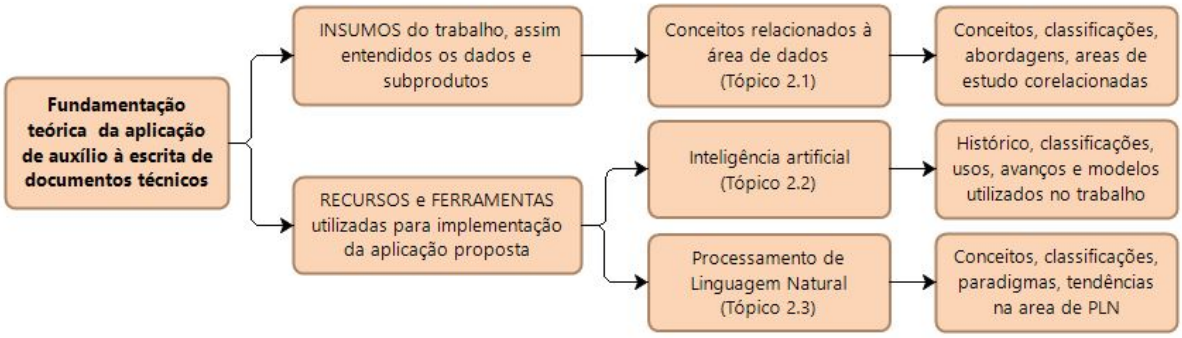
Em sequência, o Capítulo 4 versa sobre os resultados obtidos pela solução implementada, baseada em recuperação de informações de deliberações do TCU, ao que se segue o Capítulo 5 que apresenta as conclusões obtidas a partir dos esforços teóricos e práticos envolvidos no trabalho realizado, bem como as considerações acerca de possíveis aperfeiçoamentos vislumbrados e potenciais identificados para ampliação do uso da solução desenvolvida.

2 FUNDAMENTAÇÃO TEÓRICA

Neste capítulo são apresentados os fundamentos teóricos relacionados ao presente trabalho, o que se fez tangenciando conteúdos que ajudam a localizar o objeto de pesquisa no vasto campo das disciplinas envolvidas.

Neste contexto, inclui-se a apresentação de conceitos, recursos e ferramentas utilizadas no desenvolvimento do objeto deste trabalho, seguindo-se a organização representada na figura a seguir:

Figura 1 – Abordagem adotada na fundamentação teórica



Fonte: Elaboração própria

Passa-se, assim, à apresentação de conceitos relacionados à exploração de dados.

2.1 CONCEITOS RELACIONADOS À ÁNALISE DE DADOS

A famosa frase do matemático britânico Clive Humby, “Dados são o novo petróleo” (*Data is the new oil*), a cada dia se reafirma pelos esforços das companhias de tecnologia para obtenção e análise dos variados tipos de dados que nos cercam – demonstrando que, de fato, dados são insumos com inegáveis potenciais econômicos, políticos e sociais.

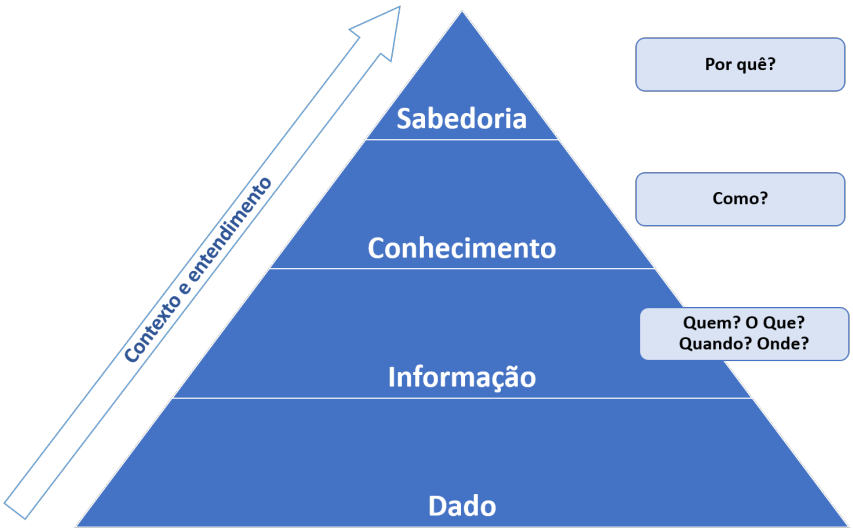
Assim como o petróleo, os dados, em seu “estado bruto”, não possuem tanta utilidade. Daí a importância dos processos de beneficiamento desses recursos. No caso do citado combustível fóssil tem-se os processos de refinamento; no caso dos dados o “beneficiamento” envolve procedimentos de seleção, limpeza, tratamento e transformações, que antecedem as técnicas analíticas.

Por outro lado, diferentemente do petróleo, os dados não são um recurso escasso. Ao contrário, atualmente dados cercam massivamente o homem contemporâneo – não

sendo exagero dizer que os dados é que estão “pastoreando” a espécie humana para a um destino incerto. A esse insumo abundante são dedicadas as próximas partes deste trabalho.

Para tanto, adota-se como organização para o presente subtópico a estrutura da conhecida Pirâmide do Conhecimento, também intitulada Hierarquia DIKW (que decorre da abreviatura, em inglês, de *Data, Information, Knowledge, Wisdom*), em português: dado, informação, conhecimento e sabedoria. Trata-se de uma estrutura de informações hierarquizadas, estudada pelo teórico organizacional norte-americano Russell L. Ackoff 1989, na qual cada camada adiciona valor à camada que a antecede:

Figura 2 – Pirâmide do conhecimento



Fonte: Elaboração própria

2.1.1 **Dados**

A palavra “dados” é originada de **data**, que em latim é o plural de *datum* (“aquilo que se dá”, segundo OECD 2008). Dados podem ser compreendidos como observações documentadas ou resultados de medição.

Noutra definição, dados são sequências de símbolos quantificados ou quantificáveis. Enfim, de forma mais concreta, considera-se dado um documento, uma informação ou um testemunho que permite chegar ao conhecimento de algo ou deduzir as consequências de um fato.

Em geral um dado isolado não é muito útil, para uma análise consistente normalmente é necessário um montante relevante de dados, que, no presente contexto, passa-se a denominar *dataset*.

2.1.1.1 Classificação dos dados quanto aos tipos primários

Existem distintas classificações de dados. Tendo por enfoque os processos de *machine learning* que servem à presente investigação, opta-se por apresentar os dados em quatro tipos: numéricos, categóricos, séries temporais e textuais.

Os **dados numéricos** (do inglês *numerical data*), também são conhecidos como dados quantitativos (*quantitative data*), são representados por meio de números, possibilitando a realização de operações matemáticas ou estatísticas – tendem a ser o tipo de dado que demandam menores esforços de pré-tratamento para os objetivos pretendidos.

Os dados numéricos são divididos em dois grupos:

(a) Contínuos – podem assumir qualquer valor dentro de um intervalo, e podem ser divididos, infinitamente, em partes; na prática são esses dados originários de medições (a exemplo de dimensões geométricas e temperatura); e

(b) Discretos – só podem assumir valores específicos e separáveis dentro de um intervalo; são valores distintos, representando itens ou ocorrências contáveis (a exemplo do número de palavras em um texto).

Os **dados categóricos** (*categorical data*) dizem respeito a características qualitativas dos elementos de um conjunto observado, como, por exemplo, as cores de um automóvel ou seu local de licenciamento. Outros exemplos são: gênero, classe social, profissão, escolaridade e tipo sanguíneo.

Dados categóricos podem assumir apenas um número limitado, e, geralmente, fixo, de valores. Apesar de ser comum em base de dados atribuir índices numéricos a esses dados, em si eles não têm valor matemático (por exemplo, não é possível calcular uma média entre cores e cidades).

As **séries temporais** (*time series*), é uma sequência de observações de uma variável ao longo do tempo, ou seja, refere-se a dados que são coletados com uma frequência temporal específica, como segundos, minutos, horas, dias, semanas, meses, bimestres e anos.

Uma das principais características associadas às séries temporais é que observações vizinhas são dependentes. Assim, a análise desse tipo de dado possibilita a detecção de tendências das variáveis observadas – o que constitui ferramenta indispensável para distintas áreas, como finanças, economia, demografia e epidemiologia.

Por último, **dados textuais** (*text data*), são os que mais interessam ao propósito do presente trabalho, estão representados na forma de palavras, frases, parágrafos que dão forma à comunicação por meio de textos em determinada língua.

Atualmente, pode-se afirmar que os dados textuais estão no centro dos interesses comerciais e acadêmicos na área da inteligência artificial, com uma ampla gama de

aplicações que vão da recuperação da informação, sumarização, análise de sentimento ou *chatbots* baseados em aplicações generativas da língua.

2.1.1.2 Classificação dos dados quanto a sua organização

Um outro critério para classificação de dados diz respeito à forma como estão organizados, o que está associado diretamente à facilidade com a qual se dar o acesso a eles. Sobre esse aspecto, são elencadas as seguintes classificações[1]:

(a) **Dados estruturados** (*Structured data*) são padronizados em um formato tabular de linhas e colunas, facilitando o armazenamento, análise e o processamento por algoritmos. São exemplos de dados estruturados aqueles armazenados, por exemplo, nos campos "nome", "endereço" e "número de telefone", em um banco de dados que pode ser acessado via linguagem SQL;

(b) **Dados semiestruturados** (*Semi-structured data*) são uma mistura entre os formatos anteriores. Embora tenha alguma organização, não se adequam aos rígidos requisitos de um banco de dados relacional. Exemplos de dados semiestruturados incluem arquivos XML, JSON e HTML – formatos que desfrutam de crescente interesse em razão do desenvolvimento de Big Data e de ferramentas e aplicações baseadas em NoSQL; e

(c) **Dados não estruturados** (*Unstructured data*) não possuem um formato de dados predefinido, representam o fenômeno da linguagem, são normalmente o tipo dados mais comumente encontrados no cotidiano, como textos da internet, relatórios técnicos e pareceres jurídicos.

2.1.2 Informação

Apesar de ser difícil definir o que é informação, pode-se assumir que a informação constitui o julgamento formado a partir de dados isolados ou agrupados, devidamente interpretados e analisados (Vieira 1995).

Noutra concepção, informação representa as inovações do conhecimento humano, em todos os ramos e atividades, e pode resultar tanto da ação mental não estruturada e informal quanto do raciocínio organizado formalmente estruturado e concatenado (Furtado 1982, apud Vieira 1995).

De acordo com Hayes 1986, informação resulta de um processo realizado com os dados ou com uma propriedade resultante destes dados. Para o autor, o referido processo, que produz a informação, pode se dar pela transmissão, seleção, organização ou análise dos dados.

[1] Informações extraídas do site "What is text mining?", disponível em urlhttps://www.ibm.com/topics/text-mining

Ainda conforme Vieira 1995, o conceito de informação, pressupõe um sistema de comunicação no qual se relacionam um emissor e um receptor. O emissor utiliza mecanismos de informação para transmitir conhecimentos que por sua vez devem ser absorvidos pelo receptor para sua utilização.

Importa anotar que, ao lidar com informações, seja para fins acadêmicos ou comerciais, deve-se ter o cuidado de observar a qual nível de classificação a informação está associada, nesse sentido elenca-se, conforme Machado 2020:

(a) Informação pública: pode vir a público sem consequências prejudiciais ao funcionamento normal da organização, neste caso a integridade não é vital;

(b) Informação interna: deve ter seu acesso restringido, embora não haja consequências prejudiciais relacionadas ao acesso ou ao uso não autorizado; a integridade deste tipo de informação é importante e precisa ser protegida, mesmo que eventuais falhas não tendam a representar impactos de grande relevância para a organização;

(c) Informação confidencial: confinada aos limites da organização, sua divulgação não autorizada ou perda pode acarretar impactos operacionais, prejuízos financeiros ou perdas de imagem ou de competitividade – essa informação deve estar submetida a rígidos controles de acesso, sua transmissão deve fazer uso de criptografia e seu descarte deve seguir procedimentos de segurança; e

(d) Informação secreta ou restrita: a segurança e a integridade deste tipo de informação são cruciais para a organização, portanto, seu acesso deve ser restrito a um número bastante reduzido de pessoas, devendo ser implementados procedimentos e políticas de segurança ainda mais rígidas do que as elencadas para a informação confidencial.

Tendo sido apresentados conceitos básicos relacionados a “dado” e “informação”, com base na pirâmide do conhecimento ilustrada na Figura 2, passa-se a comentar sobre a definição de “documento”, para o qual é associado um subtópico ainda mais específico dedicado aos documentos de natureza jurídica – visto ser esse o tipo de maior interesse para este trabalho.

2.1.3 **Documento**

Os primeiros esforços para conceituação formal de documento são atribuídos a Paul Otlet, 1934. Em sua obra, traduzida para o português como “O tratado da documentação: o livro sobre o livro: teoria e prática”, o autor elenca elementos característicos que definem um documento – termo que, na referida obra, não se restringe apenas a documentos textuais, mas, também, objetos iconográficos e audiovisuais (Souza 2013).

Assim como para os outros termos tratados neste capítulo, existem distintas definições para “documento”. Para se ter uma ideia da amplitude das concepções relacionadas ao

assunto cita-se Suzanne Briet 1951, que ampliou o conceito de documento, ao considerar que objetos colecionados em museus e animais vivos catalogados e expostos em zoológicos poderiam ser considerados documentos (Souza 2013):

> Uma estrela é um documento? Um seixo rolando em uma torrente é um documento? Um animal vivo é um documento? Não. Mas os documentos são fotografias e catálogos de estrelas, as pedras de um museu de mineralogia, os animais catalogados e expostos em um zoológico (Briet 1951), tradução nossa). (...) [Documento é] qualquer elemento concreto ou simbólico, conservado, ou registrado para fins de representar, reconstituir ou provar um fenômeno físico ou intelectual (Briet 1951, apud Souza 2013).

Buckland 1991 retoma a definição de documento como fonte de informação. Segundo o autor, os objetos só devem ser considerados como documentos quando forem percebidos nesse sentido. Assim, um objeto só poderá ser conceituado como documento quando alguém lhe atribuir a aptidão informativa.

A despeito das variadas concepções atribuídas ao termo "documento", para fins da presente investigação, "documento" será tomado como texto (escrito) que contém a informação relevante para o usuário que a busca.

Neste contexto, o conceito adotado neste trabalho aproxima-se daquele anotado por Vieira 1995, que, baseada em Pietsch 1966, define que um documento é constituído por elementos que se inter-relacionam: (i) o emissor (autor); (ii) a informação expressa em uma linguagem e fixada em um suporte; e (iii) o receptor (leitor). Tem-se, portanto, um documento como uma unidade transmissora e intermediária que utiliza a linguagem para o conhecimento de outras pessoas.

Ainda que se restrinja o conceito de documento a texto escrito, a depender dos objetivos para o qual foi produzido (ou para qual receptor é endereçado), haveria uma infinidade de possíveis classificações de documentos, como por exemplo: notícias da imprensa, propostas comerciais, pareceres técnicos ou mesmo uma lista de compra de supermercado.

Mesmo que produzidos em uma mesma língua, cada tipo de documento pode apresentar amplas variações do uso dessa língua, a depender do gênero textual e de aspectos associados ao conteúdo da mensagem que se quer transmitir – o que envolve contexto, objetivo e público-alvo.

Desta forma, faz-se necessário salientar que o enfoque do presente trabalho diz respeito especificamente a textos de natureza jurídica, o que justifica discorrer sobre algumas características relevantes deste tipo de documento.

2.1.3.1 **Documentos Jurídicos**

Essencialmente, as informações de natureza jurídica podem ser classificadas em três tipos: (i) legislação, (ii) doutrina, e (iii) jurisprudência.

A **legislação** pode ser definida como um conjunto de normas jurídicas de natureza coercitiva sobre determinado assunto, constitui a totalidade das leis de um Estado ou de determinado ramo do Direito – sendo caracterizada por (Naufel 1965, apud Vieira 1995):

(a) ser produzida apenas pelo poder estatal competente; e

(b) ser de uso público, podendo ser utilizada, coletada ou reproduzida por qualquer pessoa.

A **doutrina** é o entendimento particular sustentado por um ou vários autores especialistas sobre um aspecto legal (que normalmente não se encontra integralmente delineado pelo ordenamento jurídico).

Com base em Reale 2001, apud Souza 2013, a doutrina é uma das "molas propulsoras, e a mais racional das forças diretoras, do ordenamento jurídico". Consiste, portanto, no trabalho científico dos juristas, que criam os modelos doutrinários ou modelos dogmáticos. A Doutrina não estabelece normas, pois tem como objeto determinar no que consiste o significado das disposições produzidas pelas fontes do Direito. No entanto, a lei não poderia atingir a sua plenitude de significado ou se atualizar, se não tivesse como antecedente, o trabalho científico produzido na doutrina.

Pode-se definir **jurisprudência**, define como "a sábia interpretação e aplicação das leis a todos os casos específicos que são submetidos a julgamento. Ou seja, o hábito de interpretar e aplicar as leis a fatos concretos" (Silva 2004, apud Vieira 1995).

Para Passos e Barros (2009), apud Souza 2013, jurisprudência corresponde ao "conjunto de decisões reiteradas de juízes e tribunais sobre determinada tese jurídica, revelando o mesmo entendimento, orientando-se pelo mesmo critério e concluindo do mesmo modo", ou, em síntese, "conjunto uniforme e constante das decisões judiciais sobre casos semelhantes".

Uma pesquisa de jurisprudência pode ser orientada pelo assunto e/ou pelo tribunal em que ela foi produzida. No caso do presente trabalho, por exemplo, pode-se antecipar que para formação do repositório de documentos (que serve de fonte para a ferramenta de sugestão à escrita de documentos) foi selecionada jurisprudência com base em ambos os critérios, visto que a pesquisa foi restrita a decisões emanadas pelo TCU relacionadas a determinadas áreas de atuação do órgão – esses critérios serão melhor delimitados quando da apresentação da metodologia deste trabalho.

Em complemento, quanto às **funções dos textos jurídicos**, vale citar: (i) descritiva: consiste em relatar o fato ou experiência; (ii) explicativa: faz a ligação entre fatos e fenômenos, formulando hipóteses sobre as relações entre eles; (ii) valorativa: emite juízo de valor sobre a ação humana; e (iv) normativa: regula a ação humana por meio de disposições legais.

Quanto às **características dos textos jurídicos**, deve-se ter em mente que tais documentos não têm por objeto a teoria ou filosofia do direito, mas sim suas aplicações. Nesse sentido, vale pontuar que praticamente todos os tipos de conhecimento estão relacionados ao Direito, já que este campo abrange todas as facetas da vida, desde o nascimento do homem até depois de sua morte (Passos 1994, apud Vieira 1995).

Nessa esteira, na linguagem jurídica, destaca López *et al.* 1972, apud Vieira 1995, podem ser encontrados três tipos diferentes de termos: (i) termos estritamente jurídicos; (ii) termos técnico-científicos de outras áreas; e (iii) termos da linguagem comum. Ressalta-se que tais características devem ser consideradas no desenho da aplicações de processamento da linguagem voltados a textos jurídicos – a exemplo de ferramentas voltadas à recuperação de informações jurídicas (Passos 1994).

Outro aspecto que deve ser considerado no processamento da informação jurídica, diz respeito à longevidade. Conforme (Vaswani *et al.* 2017), não existem fórmulas matemáticas, universalmente aceitas, que possam medir a degeneração da informação. Contudo, pode-se afirmar que alguns tipos de informação se tornam obsoletos mais rapidamente do que outros, nessa categoria estão os conteúdos tecnológicos e jurídicos.

2.1.4 **Conhecimento**

Voltando-se ao terceiro nível da pirâmide ilustrada na Figura 2, o conhecimento se estabelece com base no tratamento do extrato das camadas anteriores (dados e informação). Pode-se conceituar conhecimento, como a compreensão de um conjunto de informações e maneiras como essas informações podem ser úteis para apoiar uma tarefa específica ou para chegar uma decisão.

Em complemento, segundo Burnham *et al.* 2005, apud Oliveira e Carvalho (2008): "Conhecimento é uma mistura fluída de experiência condensada, valores, informação contextual e introspecção experimentada, a qual proporciona uma estrutura para a avaliação e incorporação de novas experiências e informações".

Com base nas definições apresentadas, ressalta-se que o conhecimento, por ser derivado da informação, não é estanque, mas sim dinâmico, como um sistema vivo, que deve ser constantemente alimentado e estar exposto à interação com o meio.

2.1.5 Sabedoria

A partir do conhecimento tem-se discernimento a respeito de um assunto e adquire-se a aptidão de trabalhá-lo em conexão com outros conhecimentos já acumulados. Esse processo leva ao que se entende por sabedoria – a quarta e última camada da pirâmide do conhecimento (Figura 2). Enfim: a sabedoria possibilita a capacidade de utilizar o conhecimento acumulado de forma eficiente e tempestiva.

Cabe destacar que a tecnologia é ferramenta fundamental para os processos que envolvem extração, tratamento e análise de dados, proporcionando a transposição da informação ao conhecimento, para posterior aquisição da sabedoria de como melhor aplicá-lo.

2.2 INTELIGÊNCIA ARTIFICIAL

Este tópico tem por objetivo apresentar conceitos e fatos relacionados à inteligência artificial, ao aprendizado de máquinas e às redes neurais. Ainda que sem pretensões de esgotar os assuntos tratados, almeja-se abordar brevemente ferramentas e recursos que são úteis para a presente pesquisa.

Partindo do mais geral para o mais específico, apresenta-se uma visão global de conceitos e iniciativas que conduziram o desenvolvimento da inteligência artificial (item 2.2.1). Passa-se, então, pela abordagem de técnicas de aprendizado de máquina (item 2.2.2), para, em seguida, adentrar especificamente no aprendizado profundo (item 2.2.3), em que o escopo da explanação fica restrito a arquiteturas/modelos que estão diretamente associados à solução proposta neste trabalho.

2.2.1 Conceitos e características

Segundo o dicionário Merriam-Webster[2], Inteligência Artificial (IA), do inglês *Artificial Intelligence* (AI), é um ramo da ciência da computação que lida com a simulação do comportamento inteligente em máquinas. Essa definição tornou-se bastante difundida, bastando para comprovar isso uma rápida pesquisa do termo na internet.

Apesar de não ser incorreto localizar a inteligência artificial como um ramo da ciência da computação, deve-se sempre ter em mente que o tema envolve assuntos como aprendizado, percepção, memória e representação do conhecimento, dentre outros. Assim, trata-se de um objeto de estudo multidisciplinar de interesse crescente de áreas como matemática, estatística, ciência cognitiva, filosofia da mente, linguística computacional etc.

Cabe anotar que "inteligência" é uma palavra derivada do latim composta pela junção dos radicais *inter* ("entre", em português) e *legere*, que significa "escolher". Assim, a própria etimologia da palavra indicia que a inteligência está associada à capacidade

[2] Disponível em https://www.merriam-webster.com/dictionary/artificial%20intelligence

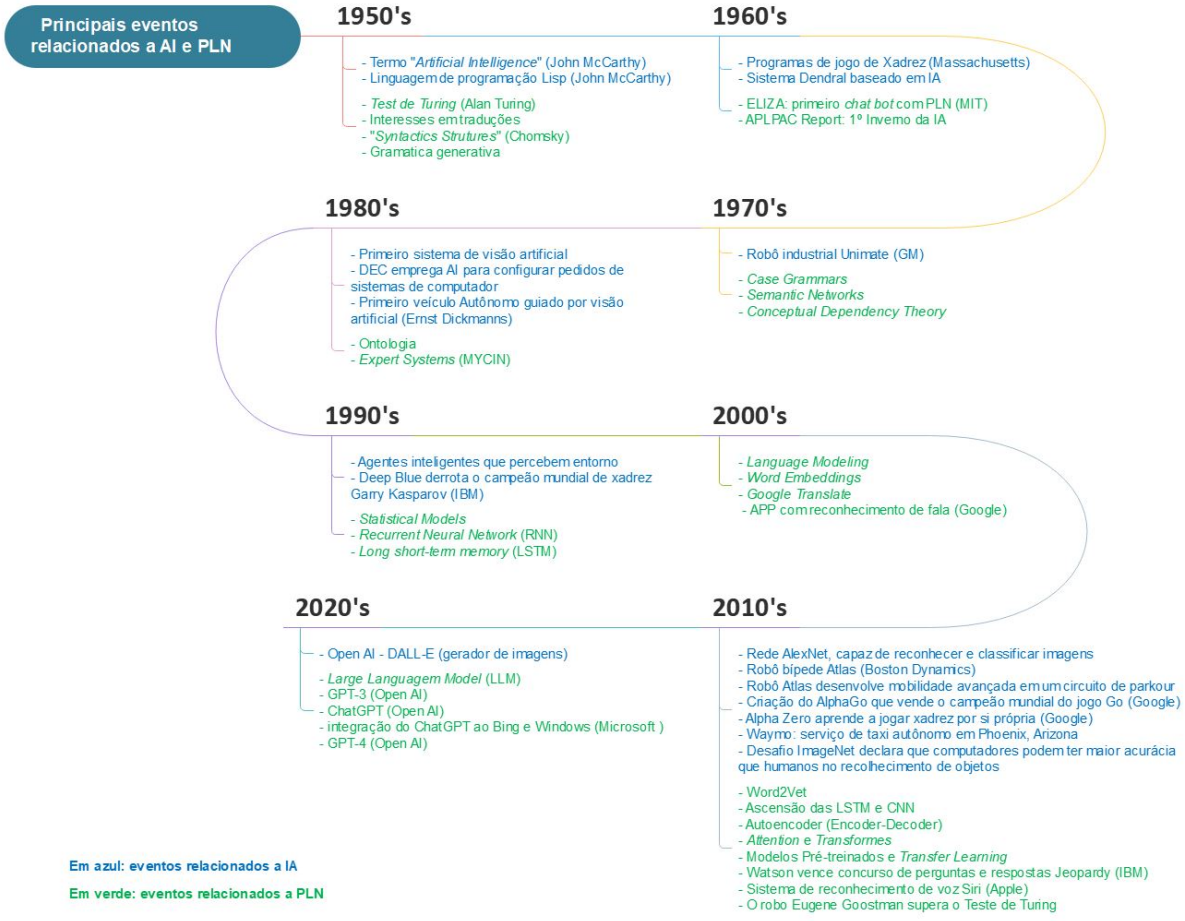
de escolha – condição essencial para resolução de problemas ou execução de tarefas (Fernandes 2003).

Desde a Grécia Antiga há discussões sobre a capacidade de objetos inanimados pensar e agir como humanos. No entanto, o conceito vigente para o termo Inteligência Artificial, remonta a 1955, quando foi proposto por John McCARTHY (McCarthy *et al.* 2006) que aspectos da inteligência humana poderiam ser simulados por máquinas (Vieira 1995).

Neste contexto, a inteligência artificial dedica-se ao estudo de paradigmas de análise de dados que não são explicitamente programados. Noutras palavras, a saída de um algoritmo de inteligência artificial não é resultado lógico e determinístico dos códigos utilizados na linguagem de programação.

Ainda que não seja o intuito da presente pesquisa discorrer detidamente sobre os numerosos eventos significativos associados à inteligência artificial, são pontuados na Figura 3 alguns dos principais avanços relacionados a essa área (organizados de forma esquemática por década):

Figura 3 – Principais fatos, por década, relacionados às áreas de IA e PLN



Fonte: Elaboração própria

Para a presente pesquisa, merece destaque os avanços ocorridos na década de 2010, dentre os quais pode-se citar o Word2Vet e arquitetura *Transformers* (a qual se dedica subtópico específico neste trabalho). Essas inovações propagaram a revolução iniciada pelos *word embeddings* (2000's) na área de processamento de linguagem natural – assuntos também retomados mais adiante.

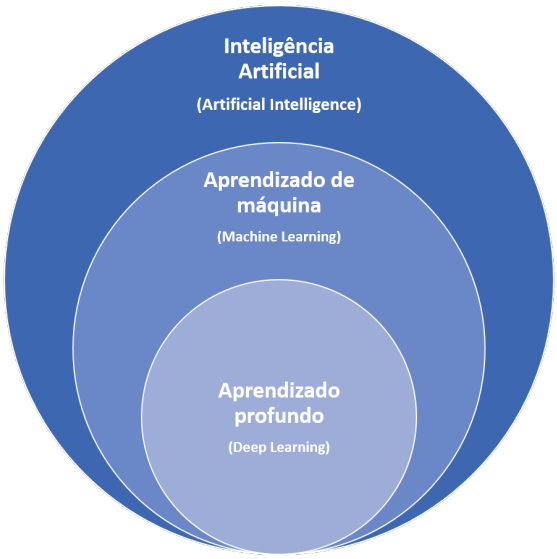
Atualmente, a inteligência artificial conquistou distintos espaços de forma definitiva, dada sua incontestável eficácia na resolução de problemas relacionados às diversas áreas em que é aplicada.

O uso da IA está se difundindo até mesmo em atividades culturais e recreativas, tais como geração de textos, músicas, imagens e filmes – avançando, sem grandes dificuldades, por campos que, até pouco tempo atrás, acreditava-se serem privativos da criatividade humana – o que evidencia a necessidade de abordagens multidisciplinares sobre o tema.

Nessa esteira, não se poderia se abster de anotar que, em novembro de 2022, o lançamento, pela empresa OpenAI, do *chatbot* ChatGPT (do inglês: *Chat Generative Pre-trained Transformer*) causou grande alvoroço por sua notória capacidade de responder a praticamente qualquer pergunta ou solicitação em língua natural, incluindo o português. Conforme anota Caseli e Nunes, 2023, além de surpreender pelo seu desempenho linguístico, o ChatGPT acendeu um sinal de alerta para a comunidade de IA e PLN, bem como para vários setores da sociedade, trazendo novas e variadas questões éticas que atualmente são objeto de grande atenção.

Dentro do campo de estudo da inteligência artificial, encontra-se o aprendizado de máquina, que, por sua vez, envolve o aprendizado profundo – compondo uma relação que se costuma representar por ilustrações similares à Figura 4.

Figura 4 – Áreas de estudo AI e Aprendizado de máquina



Fonte: Elaboração própria

Com base na Figura 4 – adotando-se o sentido de fora para o centro como estratégia de abordagem do conteúdo a ser apresentado –, dedica-se o subtópico a seguir aos comentários sobre aprendizagem de máquina.

2.2.2 **Aprendizado de Máquina**

O objetivo do aprendizado de máquina (*machine learning*) é criar modelos computacionais que, treinados com grandes volumes de dados, tornam-se capazes de prever resultados, tendências e padrões (Barnes 2015). Nessa esteira, pode-se afirmar que o aumento da produção, coleta e disponibilidade de dados, propiciado pela internet e pela difusão das redes sociais, foi um dos elementos fundamentais para o desenvolvimento da área de aprendizado de máquina.

Quanto aos objetivos das tarefas de aprendizado de máquinas, tem-se a divisão em dois grandes grupos:

(a) **tarefas descritivas**, que consistem basicamente em explorar ou descrever um conjunto de dados, objetos, ou exemplos. Nessa categoria estão as tarefas de categorizar os elementos de conjunto ou agrupá-los conforme as semelhanças de seus atributos; e

(b) **tarefas preditivas**, voltadas a encontrar uma função, modelo ou hipótese que pode ser utilizada para predizer um atributo de um elemento desconhecido ao conjunto; como por exemplo, prever um valor de um imóvel, de acordo com suas especificações, partindo de uma base de dados utilizada para treinamento do modelo.

Uma outra classificação, que diz respeito, não aos objetivos das tarefas executadas, mas primordialmente às propriedades dos dados de entrada disponíveis, divide o aprendizado em (i) supervisionado; (ii) semissupervisionado (iii) não-supervisionado; e (iv) por reforço.

O **aprendizado supervisionado** (*supervised learning*, em inglês) ocorre quando o modelo é treinado a partir de dados que contém os resultados reais da classificação ("dados rotulados"). Esses valores servem para avaliação (ou "supervisão") dos resultados das predições executadas – permitindo o ajuste das previsões posteriores com base nos erros identificados.

Esse modelo presta-se essencialmente a tarefas de classificação, em que, por meio da identificação de padrões associados a cada categoria, o algoritmo tenta predizer a classe de um novo elemento avaliado.

O **aprendizado semissupervisionado** (*semi-supervised learning*, em inglês) é utilizado para casos em que apenas parte dos dados de *input* fornecidos é rotulada. A rigor, a dependência, ao menos parcial, de dados rotulados faz com que esse tipo de técnica

possa ser compreendido como tipo específico de aprendizado supervisionado[3].

Anota-se que o aprendizado semissupervisionado, bem como o não-supervisionado (comentado em seguida), representam alternativas eficientes em muitas situações; visto que pode ser demorado e dispendioso, ou mesmo inviável, depender do conhecimento amplo de determinado domínio para rotular dados (como requer o aprendizado supervisionado).

No **aprendizado não-supervisionado** (*unsupervised learning*, em inglês) não existem resultados pré-rotulados para o modelo utilizar como referência para aprender.

Uma técnica associada a esse tipo de aprendizado é o *clustering*, que, basicamente, pode ser descrito como o processo de agrupar e categorizar conjuntos de dados. Nesse caso, o modelo tem a tarefa de encontrar semelhanças e diferenças entre os dados, para assim propor a separação dos elementos em categorias. Esse tipo de modelo é capaz de detectar padrões desconhecidos nos dados, que, não raras vezes, são de difícil percepção até por um humano.

Cabe observar que no aprendizado supervisionado, medir o desempenho pode ser verificado de forma simples e objetiva – por meio da divisão dos dados em treino e teste, é possível avaliar os resultados preditos com base nos rótulos pré-definidos. No entanto, no aprendizado não supervisionado, muitas vezes, nem sequer há uma "resposta correta" que possa servir para avaliar o modelo de forma objetiva.

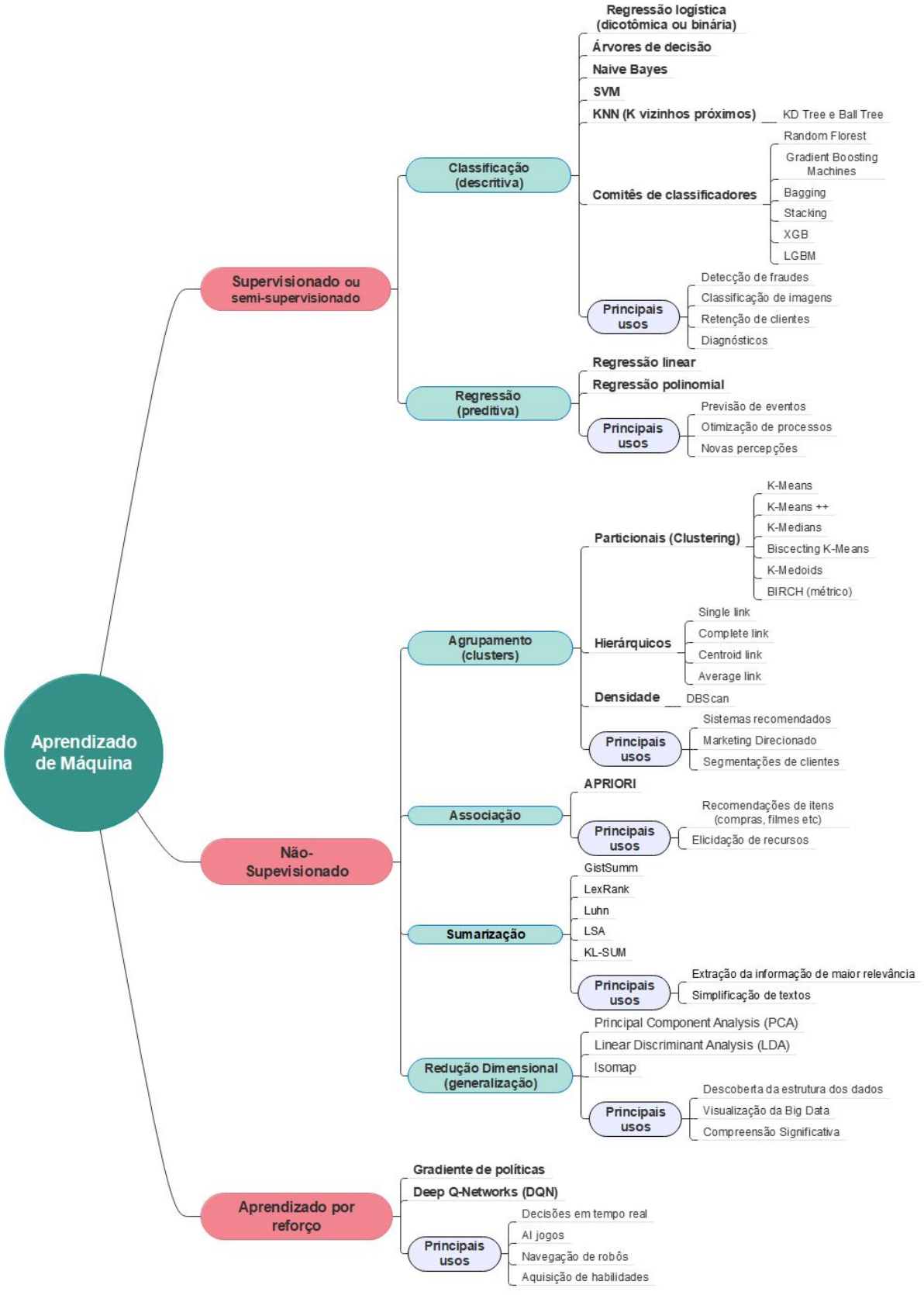
No **Aprendizado por reforço** (do inglês, *Reinforcement Learning*) um agente (entidade real, como um robô, ou simulada por software) é treinado para tomar sequências de decisões, pelas quais recebe recompensas (ou penalidades). O objetivo é que o agente aprenda a escolher as ações que maximizam as recompensas ao longo do tempo. Esse tipo de aprendizado é amplamente aplicado na robótica e em jogos.

Importa observar que, apesar de basear-se em técnicas bastante específicas, em certo aspecto, o aprendizado por reforço aproxima-se do aprendizado supervisionado na medida em que para ambos existem informações sobre a resposta (ou ação) correta para treinar e avaliar o desempenho dos resultados fornecidos pelo modelo.

Em arremate, são apresentados na Figura 5 a seguir algoritmos, modelos e técnicas, que servem para localizar, em forma de esquema, não rígido e tampouco exaustivo, alguns dos principais objetos de estudo abrangidos pela área de aprendizado de máquina (não cabendo ao escopo do presente trabalho, entretanto, o detalhamento de características de tais elementos).

[3] Informações extraídas do site "O que é aprendizado supervisionado?", disponível em https://www.ibm.com/br-pt/topics/supervised-learning

Figura 5 – Classificação de aprendizado de máquina e principais técnicas e usos



Fonte: Elaboração própria

Para organização dos tipos de técnicas elencados na Figura 5 adotou-se aquela tarefa/função que mais comumente serve de base para a classificação desses algoritmos. Com isso, insta reconhecer que o esquema apresentado, comete imprecisão ao não levar em consideração que uma mesma técnica pode ser utilizada em tarefas de natureza distintas – como é o caso, por exemplo, de árvores de decisão, regressão linear e SVM, que, com as devidas adaptações, podem realizar tanto tarefas preditivas como descritivas.

Dando sequência à explanação que visa a progressiva aproximação dos conceitos gerais ao objeto particular da presente pesquisa, passa-se ao tópico dedicado especificamente ao aprendizado profundo, no qual encontra-se os comentários sobre a arquitetura e modelo neurais utilizados para aplicação proposta neste trabalho.

2.2.3 **Aprendizado profundo**

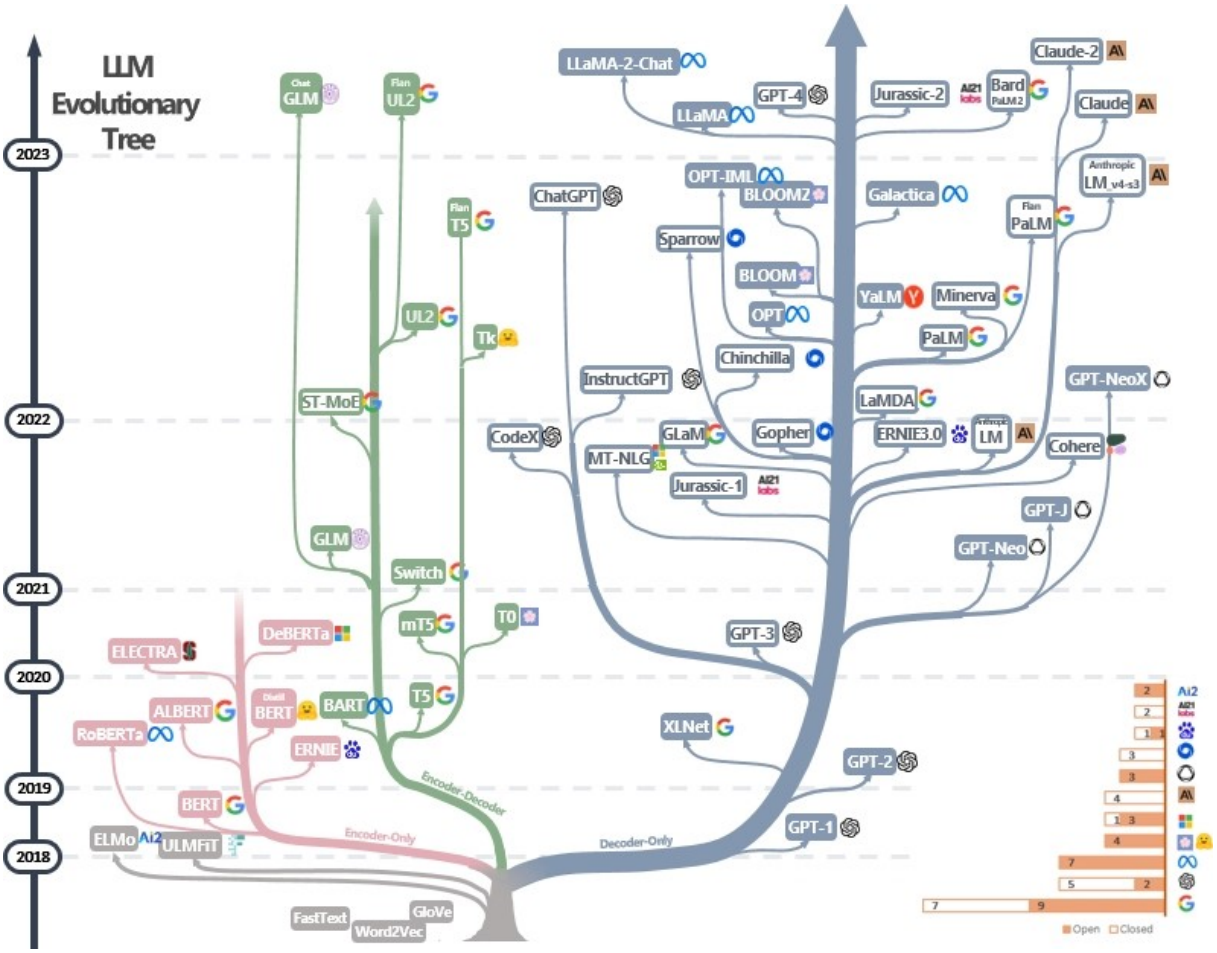
O aprendizado profundo (do inglês *Deep Learning*) diz respeito a modelos computacionais compostos por múltiplas camadas de processamento, que permitem aprender sobre dados que possuem muitos níveis de abstração. Frise-se que o termo "profundo" se refere justamente ao número de camadas de redes neurais; caracterizando um aprendizado hierárquico, no qual o modelo extrai conhecimento a partir dos dados de entrada (Eddison 2017).

Em linhas gerais, no processo em comento, os dados são alimentados através de redes neurais, que realizam uma série de verificações binárias de verdadeiro e falso, ou extraem valores numéricos. Cada bit de dados que passa através dessas redes neurais é classificado conforme as respostas processadas anteriormente (Eddison 2017).

No aprendizado profundo, devido à grande complexidade inerente às arquiteturas compostas por diversas camadas de processamento, em geral, não é possível alcançar um razoável nível de explicabilidade acerca do processo percorrido para chegar a determinada resposta – o que a depender dos objetivos pode se tornar uma questão relevante, a exemplo, do direito do cliente em obter resposta do motivo para ter um empréstimo bancário rejeitado por um algoritmo baseado em redes neurais.

A eficiência de algoritmos baseados em aprendizado profundo envolve a disponibilidade de vastos conjuntos de dados. Nesse cenário, merecem destaque os modelos conhecidos por *Large Language Models* (LLM), que usam algoritmos de aprendizado profundo para processar e entender a linguagem natural. Na figura a seguir apresenta-se uma árvore evolutiva desta área (Yang *et al.* 2023):

Figura 6 – Árvore de evolução dos principais modelos baseados em LLMs



Fonte: (Yang *et al.* 2023)

Na Figura 6, modelos de código aberto são representados por quadrados sólidos, ao passo que os de código fechado estão associados aos quadrados vazados. No canto inferior à direita, o gráfico de barras empilhadas apresenta quantidade de modelos associada a cada uma das empresas/instituições desenvolvedoras.

Ainda com base na Figura 6, verifica-se que com exceção dos poucos modelos representados dos dois ramos menores (mais baixos à esquerda), todos os demais são baseados na arquitetura neural denominada *Transformers* – diferenciando-se por fazerem uso dos módulos da citada arquitetura: codificador (cor rosa); decodificador (verde); ou ambos (cinza-azulado).

Dada a demonstrada difusão do uso dos *Transformers*, passa-se a tratar da referida arquitetura de rede neural; para, em sequência, abordar especificamente um dos modelos baseados nela (a qual serviu de base a este trabalho): o *Bidirectional Encoder Representations from Transformers* (BERT).

2.2.3.1 **Transformers**

Em um artigo de 2017, intitulado "*Attention Is All You Need*" (Vaswani *et al.* 2017), escrito por membros do Google Brain (fundido ao Google DeepMind, em 2023) e do Google Research, foi proposta uma nova arquitetura de aprendizagem profunda denominada *Transformer*.

No citado artigo, para demonstração da utilização do modelo baseado em *Transformers* foram utilizadas tarefas de traduções do inglês para francês e do inglês para alemão. Os resultados obtidos reconhecidamente desbancaram as arquiteturas anteriores (que eram baseadas especialmente em redes neurais recorrentes).

Desde então, variações do modelo inicial baseado nos *Transformers*, se tornaram o novo estado da arte em tarefas de processamento de linguagem natural (área a qual é destinado um tópico específico mais adiante neste trabalho) – o que decorre, especialmente, pela capacidade destes modelos de considerar o contexto em que cada palavra está inserida e com isso apreender seu significado.

Os *Transformers* são a base (e compõem os nomes), por exemplo, dos modelos *Generative Pre-Trained Transformer* (GPT) e *Bidirectional Encoder Representations from Transformers* (BERT) e suas variantes. Atualmente, os *Transformers* também são amplamente adotados para treinar modelos robustos de linguagem, *Large Language Model* (LLM), bem como grandes conjuntos de dados linguísticos, como o *corpus* da Wikipédia.

Seu desempenho e versatilidade fez com que a arquitetura *Transformers* passasse a ser utilizada também em aplicações que lidam com series temporais, visão computacional, áudio e processamento multimodal multimídia.

O grande salto de desempenho associado aos *Transformers* deve-se ao uso do mecanismo ***attention*** (daí o sugestivo título do artigo na qual a arquitetura fora apresentada). Traduzido para o português como "mecanismo de atenção" ou "mecanismo de autoatenção", permite capturar o contexto em que determinada palavra está inserida, por meio da percepção de dependências complexas entre palavras (mesmo quando distantes no texto).

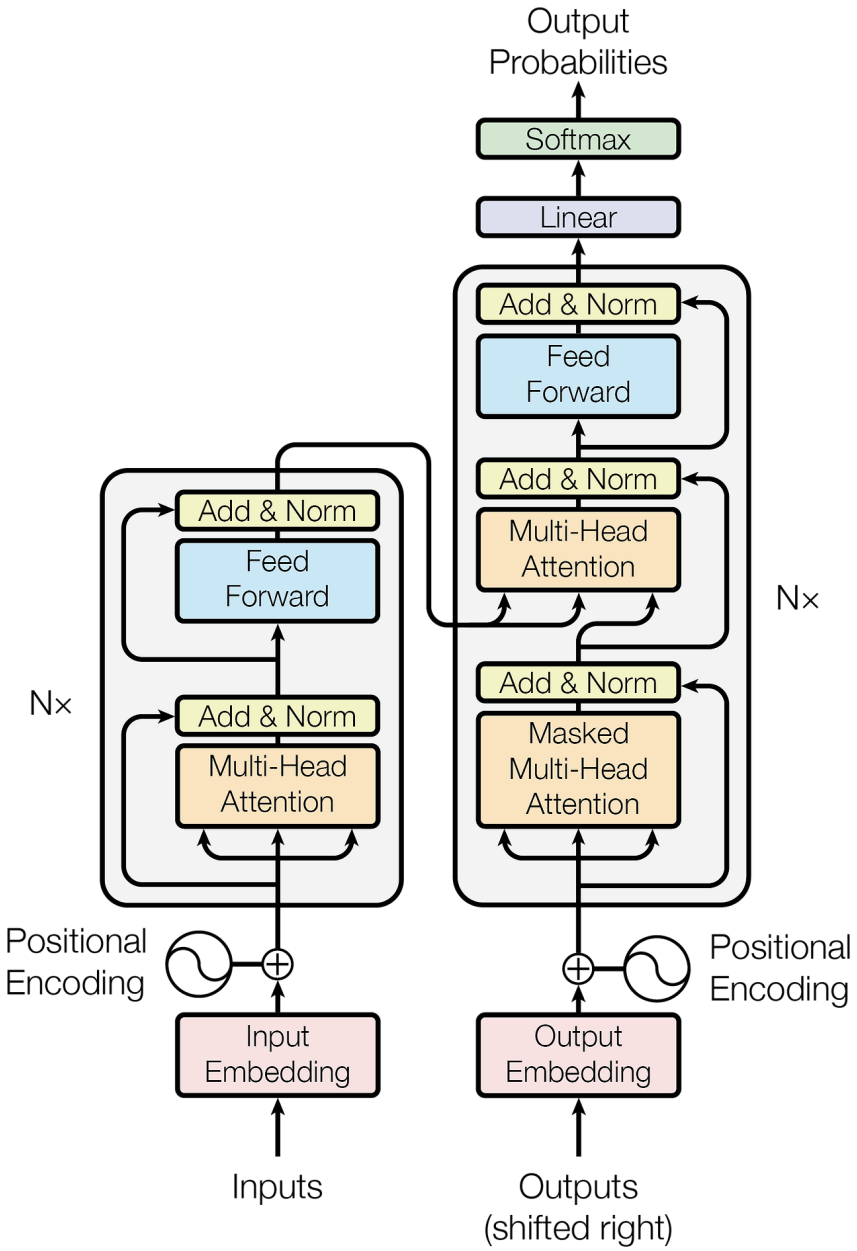
Com isso, os *transformes* podem subsidiar modelos capazes de reconhecer distintos significados da mesma palavra inserida em contextos diferentes; sendo possível, por exemplo, distinguir, a partir do contexto de uma sentença, se a palavra "manga" foi utilizada para significar fruta ou parte da vestimenta. Essa característica se demonstra especialmente relevante na tradução de línguas e no processamento de palavras novas (fora do vocabulário).

O mecanismo de atenção realiza uma operação "palavra a palavra" para descobrir como cada palavra está relacionada a todas as outras em uma sequência (Rothman, 2021). Assim, o referido mecanismo se coloca em substituição à estratégia de recorrência, em que se baseia a *Recurrent Neural Network* (RNN) – em que conjuntos de dados são processados

sequencialmente, exigindo um número crescente de operações à medida que a distância entre duas palavras se distende.

Ainda que não seja o objetivo deste trabalho esmiuçar a **arquitetura *Transformers***, vale tecer mais alguns comentários sobre o esquema original do modelo proposto por Vaswani *et al.* 2017, partindo da figura 7.

Figura 7 – Arquitetura original dos *Transformers*



Fonte: Vaswani *et al.* 2017

Destaca-se que a arquitetura *Transformers* caracteriza uma estrutura denominada *autoencoder*, constituída de (i) uma parte denominada ***Encoder*** que compacta a entrada em uma representação eficiente do espaço latente e (ii) outra parte que reconstitui a

entrada da representação do espaço latente, denominada ***Decoder***. O modelo original do *Transformer* possui uma pilha de seis camadas de *encoders* e outra pilha de *decoders*, também de seis camadas.

Com base na figura apresentada, observa-se que as entrada dos transformes recebem um *embedding* de entrada que é somado a um *embedding* de posição, com vista a inserir na representação vetorial uma identificação da posição de cada token na sequência de elementos de entrada.

As entradas fornecidas ao codificador passam através de subcamada *Attention* e da subcamada *FeedForward Network* (FFN). A saída do *encoder* se somam ao aos *outputs embbedings* que vão para o lado do decodificador. O elemento ***multi-head attention*** (característico da arquitetura em comento), em essência, tem por função encontrar correlações entre os pares de palavras de uma frase.

Do lado do decodificador também estão presentes as comentadas subcamadas de atenção (*Attention*). Cabendo observar que no decodificador existe uma dessas subcamadas mascarada (*masked*), a qual é repassada uma tarefa de prever palavras que não foram comunicadas a esse módulo, isso força o *Transformers* a aprender (Rothman 2021).

Anota-se que sempre após uma subcamada, seja do codificador ou do decodificador, aplica-se ao *embeddings* as funções (i) de adição de resíduo (o vetor entrada da subcamada é somada ao vetor de saída) e (ii) normalização.

Por último, cabe o registro que modelos baseados em *Transformers* constam de distintas bibliotecas especializadas em aprendizado de máquina, como TensorFlow e PyTorch. Anota-se ainda que a plataforma comunitária Hugging Face fornece vasta documentação, além de arquiteturas e modelos pré-treinados baseados nos *Transformers*[4].

2.2.3.2 Modelo BERT

Em novembro de 2018, o Google lançou, em código aberto, um modelo de rede neural voltada para o processamento de linguagem denominado *Bidirectional Encoder Representations from Transformers* (BERT). Em outubro de 2019, o Google anunciou a utilização do modelo BERT no algoritmo no seu site de buscas.

No artigo inaugural desta tecnologia, intitulado *"BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding"*(Devlin *et al.* 2018), foram realçadas as características desse modelo, alcunhado como "primeira representação de linguagem não supervisionada profundamente bidirecional".

Como característica principal do modelo BERT tem-se sua elevada capacidade de apreender o contexto completo de uma palavra em uma sentença, com base nos termos que

[4] Disponível em https://huggingface.co/docs/Transformers/index

vêm antes e depois e as relações entre eles – de forma análoga ao próprio comportamento humano quando se depara com dificuldades para entender uma frase.

O BERT é categorizado, portanto, como um **modelo contextual**. Sobre esse assunto, vale anotar que as representações de linguagem pré-treinadas podem ser livres de contexto ou contextuais.

Modelos livres de contexto, como Word2vec (Mikolov *et al.* 2013) ou GloVe (Pennington *et al.* 2014), geram uma única representação para cada palavra dos seus vocabulários. Por exemplo, nesses modelos, a palavra "banco" teria a mesma representação vetorial para os significados associados à conta bancária e ao assento de praça.

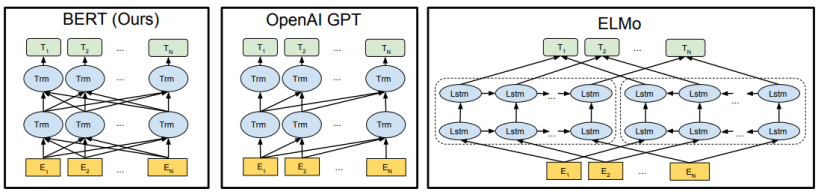
Em vez disso, os modelos contextuais podem gerar várias representações para uma mesma palavra, com base nas outras palavras da frase – capturando assim distintos contextos e, consequentemente, variados significados de uma determinada palavra.

Conforme antecipado, as redes BERTs são baseadas na arquitetura ***Transformers*** (que igualmente é fruto das pesquisas da Google). No entanto, o modelo BERT faz uso apenas da camada *encoder* do modelo *Transformers* (representado na Figura 7).

A **bidirecionalidade**, que caracteriza o modelo em comento, é obtida, pelo fato, de as frases fornecidas à camada *encoder* serem processadas tanto em relação aos dados à esquerda quanto à direita. Dessa forma, durante o treinamento, uma frase (representada por um vetor *embedding*) é "lida" tanto em direção aos dados mais à direita quanto em direção oposta, porém em um mesmo centro de processamento (Araújo *et al.* 2022).

O artigo original (Devlin *et al.* 2018), ilustra a estrutura do modelo BERT, destacando as interconexões entre os *Transformers*, em comparação com outras tecnologias utilizadas em tarefas de processamento de linguagem já disponíveis à época:

Figura 8 – Ilustração comparativa de modelos de PLN



Fonte: Devlin *et al.* 2018

O BERT definiu um novo estado da arte com seu desempenho em diversas tarefas, incluindo resposta a questões formuladas em língua natural, classificação de sentenças e regressão de pares de sentenças. Esses resultados estimularam a criação de numerosos

modelos descendentes do BERT, como o RoBERTa[5] (Liu *et al.* 2019) e DistilBERT[6] (Sanh *et al.* 2020).

No entanto, o modelo BERT e suas variações se demonstraram ineficientes para tarefas de regressão mais exigentes, devido a uma grande quantidade de combinações possíveis a serem processadas. Constatou-se ainda que a concepção do BERT não alcançava resultados desejados em pesquisas de similaridade semântica e em tarefas não supervisionadas, como agrupamento.

Para combater essas deficiências, foi proposta uma importante variação do modelo denominada *Sentence-BERT* – que interessa especialmente à solução proposta no presente trabalho, justificando que se discorra mais detidamente sobre alguns dos seus principais aspectos.

2.2.3.3 ***Sentence-BERT***

O trabalho intitulado *"Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks"* de Reimers e Gurevych, 2019, é considerado seminal na apresentação das características e aplicações deste modelo.

No citado artigo é reconhecida a importância de publicações anteriores sobre as redes BERT, como "BERT" (Devlin *et al.* 2018) e "RoBERTa" (Liu *et al.* 2019), que representaram um relevante avanço nos estudos de tarefas de regressão de pares de frases, como similaridade textual semântica, do inglês *Semantic Textual Similarity* (STS).

Em sequência, o artigo parte de uma tarefa de identificação do par de sentenças mais semelhante em uma coleção de 10.000 sentenças. Utilizando o modelo BERT original a referida tarefa envolveria um alto custo computacional e, conforme o citado artigo, demandaria aproximadamente 65 horas de processamento em um equipamento moderno da época (Reimers e Gurevych, 2019).

Segundo os autores do artigo (Reimers e Gurevych, 2019), a mesma tarefa poderia ser resolvida pelo *Sentence-BERT* em apenas 5 segundos, fazendo uso do cálculo de similaridade de cosseno (assunto objeto de subtópico específico mais adiante neste trabalho). Na figura 9 são apresentadas duas configurações do modelo *Sentence-BERT*, propostos no artigo inaugural em comento:

[5] Disponível em https://huggingface.co/docs/Transformers/model_doc/roberta
[6] Disponível em https://huggingface.co/docs/Transformers/model_doc/distilbert

Figura 9 – Configurações de uso do SBERT

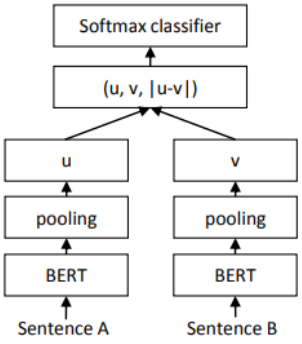


Figure 1: SBERT architecture with classification objective function, e.g., for fine-tuning on SNLI dataset. The two BERT networks have tied weights (siamese network structure).

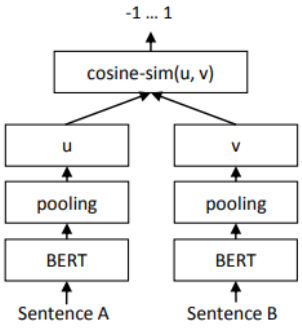


Figure 2: SBERT architecture at inference, for example, to compute similarity scores. This architecture is also used with the regression objective function.

Fonte: Reimers e Gurevych 2019

A configuração ilustrada à esquerda da Figura 9 é voltada para tarefas de classificação. Já a configuração ilustrada à direita, permite ao *Sentence-BERT* (ou SBERT) encontrar sentenças semanticamente próximas utilizando medidas de similaridade como a função cosseno ou a *Manhattan/Euclidian Distance* (Reimers e Gurevych, 2019).

Os cálculos das supracitadas medidas alcançam performance extremamente eficiente, o que justificou o uso do *Sentence-BERT* para busca de similaridade semântica, na qual se baseia a ferramenta de recuperação de informações objeto da presente pesquisa.

2.3 PROCESSAMENTO DE LINGUAGEM NATURAL

O Processamento de Linguagem Natural (PLN), do inglês *Natural Language Processing* (NLP), que evoluiu da linguística computacional, utiliza métodos de diversas disciplinas, como ciência da computação, inteligência artificial, linguística e ciência de dados, para permitir que os computadores interajam com a linguagem humana nas formas escrita ou verbal.

Conforme Caseli *et al.* 2023, o Processamento de Linguagem Natural pode ser dividido em duas grandes subáreas de pesquisa:

(a) **Interpretação (ou Compreensão) de Linguagem Natural**, do inglês, *Natural Language Understanding* (NLU): que diz respeito ao processamento com vistas à análise e à interpretação da língua; entendendo-se (i) análise como a segmentação e classificação dos componentes linguísticos, como, por exemplo, palavras e suas classes morfológicas e gramaticais, seus traços semânticos ou ontológicos etc.; e (ii) interpretação como tentativa de apreender significados construídos pelo ser humano; e

(b) **Geração de Linguagem Natural**, do inglês, *Natural Language Generation* (NLG), para qual os objetivos estão associados à produção de conteúdo com significado em linguagem humana, como por exemplo a geração de respostas ao usuário dos *chatbots*

Anota-se que o Processamento de Linguagem Natural (PLN) surgiu praticamente ao mesmo tempo que os computadores, por volta da década de 1940, já que a tradução automática entre línguas foi um dos primeiros problemas submetidos aos primeiros computadores (Caseli e Nunes, 2023).

Alguns dos mais relevantes fatos relacionados ao processamento de linguagem natural constam da linha do tempo objeto da Figura 3. Em acréscimo, sem pretensões de apresentar uma cronologia definitiva, pode-se organizar as fases de desenvolvimento do PLN nos seguintes períodos (tendo-se em mente que os conceitos mais relevantes para o propósito deste trabalho serão objeto de comentários mais adiante):

(a) **1940-1956**: fundação da área, com trabalhos voltados a máquinas de estados finitos, gramáticas, e modelos probabilísticos;

(b) **1957-1970**: os estudos realizados voltam-se principalmente para as abordagens simbolista e estatística; surgem, nessa época, os primeiros *corpus on-line*;

(c) **1970-1983**: tem-se a abordagem estocástica; e pesquisas com enfoque em interpretação de texto e na análise do discurso;

(d) **1983-1993**: predominam os modelos baseados no empirismo (abordagem estatística) e os interesses por tarefas de avaliação e de geração de textos;

(e) **2000-2012**: os interesses e investimentos voltam-se ao aprendizado de máquina (supervisionado, semissupervisionado, não supervisionado e aprendizado sem fim); associam-se também a este período o surgimento das competições de desempenho e dos grandes conjuntos de dados.

(f) **2013 em diante**: dominância da utilização de redes neurais profundas (convolução, recorrência e mecanismo de atenção), da modelagem distribucional, e dos grandes modelos de língua; tem-se ainda, mais recentemente, a popularização de ferramentas baseada em geração de textos (como o ChatGPT), com impactos em diversos aspectos da sociedade.

Atualmente, especialmente em função do difundido uso de celulares, o PLN permeia o cotidiano de grande parte da população – subsidiando a execução de ampla variedade de tarefas, tais como : (i) tradução automática; (ii) correção ortográfica ou gramatical; (iii) sistemas de classificação textual; (iv) sumarização automática de documentos (v) sistemas

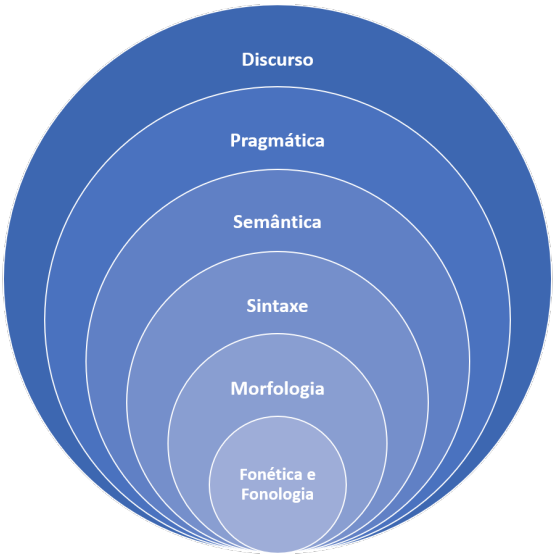
de auxílio à escrita; (vi) sistemas de recuperação de informação; (vii) sistemas de detecção de fake news; e (viii) assistentes virtuais e *chatbots*;

2.3.1 **Níveis de processamento linguístico**

Abordar aspectos da linguagem e suas variadas dimensões, mesmo que em nível elementar, exigiria dispêndios de esforços que superaria sobremaneira os objetivos deste trabalho. No entanto, faz-se necessário comentar brevemente sobre conceitos e definições associados às camadas de análise da linguagem (também chamados de "modos" ou "níveis de processamento linguístico"), visto que o assunto permeia o desenvolvimento de outras partes da presente pesquisa.

Neste contexto, compreende-se como camadas de linguagem as suas subáreas de estudo representadas conforme a Figura 10:

Figura 10 – Subáreas de estudo da linguagem



Fonte: Elaboração própria

No núcleo da estrutura da linguagem, os sons e sua organização são estudados pela **fonética e fonologia**. Línguas faladas e línguas de sinais contêm um sistema fonológico que rege a forma como os sons ou os símbolos visuais são articulados a fim de formar as sequências conhecidas como palavras ou morfemas. Assim, envolvendo a estrutura sonora, temos o estudo de como os morfemas se organizam para formar palavras, que é objeto de estudo da **morfologia**.

Envolvendo a morfologia, temos o estudo de como as palavras se organizam em estruturas para formar sintagmas e orações, caracterizando o objeto de estudo da **sintaxe** (Caseli *et al.* 2023). Línguas escritas usam símbolos visuais para representar os sons das

línguas faladas, mas elas ainda necessitam de regras sintáticas que governam a produção de sentido a partir das sequências das palavras.

No círculo que envolve a sintaxe, temos a **semântica**, que estuda o significado de palavras e frases A semântica de uma gramática é o estudo da viabilização de relacionamento entre os constituintes da estrutura gramatical mínima – ordenada de acordo com uma sintaxe.

Pode-se dizer que uma frase é compreensível se associar um significado à sua estrutura morfossintática. Nessa esteira, a semântica permite descrever os significados baseados nas palavras ou grupos de palavras e analisar que influência pode ter as categorias gramaticais e suas combinações no significado de uma sentença (Vieira 1995).

Os estudos na área da semântica não se restringem apenas a significados gerais e conceituais, nem mesmo as relações entre esses significados e a realidade dos falantes – deve-se ter em consideração, por exemplo, que em função de convenções estabelecidas por uma comunidade pode ser atribuído um ou vários significados a uma mesma palavra (Dijk 1987, apud Vieira 1995).

Acima da camada da semântica está a **pragmática** que se volta à linguagem em relação a seu uso contextual e à pessoa que a utiliza (aspectos assim considerados "fora do texto"). Noutras palavras, a pragmática enfoca como as orações são utilizadas na interação para fins comunicativos. Conforme (Vieira 1995), a tarefa da pragmática é descrever as condições (ou contexto) nas quais os enunciados são aceitáveis, apropriados ou oportunos no processo comunicativo em que se expressa o falante.

Por último, o **discurso** é uma denominação que abrange os estudos com foco no texto como um todo, podendo se referir à análise (i) das relações entre frases ou partes de um texto, ou (ii) das partes constituintes da estrutura de um texto (a exemplo da análise da articulação da "introdução", "desenvolvimento" e "conclusão" de uma redação).

2.3.2 Paradigmas de PLN

Considerando o tipo de método adotado para a análise dos dados textuais, costuma-se separar as abordagens em: (i) simbolista; (ii) estatística (empírica); e (iii) neural (conexionista). Pode-se considerar ainda a existência de aplicações baseadas em uma abordagem híbrida, combinando características das abordagens elencadas.

A **abordagem simbólica (ou paradigma simbólico)**, bastante estudado até a década de 1980, tem como base regras bem definidas e estruturadas de linguística. Basicamente, o simbolismo propunha que todo conhecimento sobre a língua seria expresso explicitamente em formalismos como léxicos, regras, linguagens lógicas etc., ou seja, formas compreensíveis ao humano (Caseli *et al.* 2023).

Neste contexto, poderiam ser definidos algoritmos para processamentos de lin-

guagem simples. Por exemplo, pode-se escrever regras que determinem que existe a concordância entre o gênero gramatical atribuído a um substantivo e o gênero atribuído ao adjetivo que o acompanha – assim a construção "abacaxi maduro" poderia ser classificada, por uma aplicação, como alinhada ao critério estabelecido; ao passo que "abacaxi madura" teria diferente tratamento (Caseli *et al.* 2023).

No início dos anos 1990, ganha interesse a chamada **abordagem empírica (ou paradigma estatístico)** que se baseia no próprio texto ou fala para realizar as suas deduções de interpretação – por meio de modelos matemáticos que dispensam o emprego das regras linguísticas (que davam suporte à concepção simbolista).

A ascensão dessa abordagem estatística, que desfrutou de grande atenção até a década de 2010, deve-se diretamente ao aumento da capacidade de memória e de processamento dos computadores, que possibilitaram o desenvolvimento de diversos algoritmos de aprendizado de máquina.

Conforme registra Caseli *et al.* 2023, grandes conjuntos de textos (também chamados de *corpus*) passaram a ser usados como fonte de conhecimento para "ensinar" as máquinas. Fenômenos como a concordância entre substantivo e adjetivo passaram a ser aprendidos a partir de exemplos obtidos dos *corpus* utilizados.

A língua passou a ser representada em modelos probabilísticos aprendidos a partir da frequência de ocorrência dos termos constituintes dos textos. Nesse cenário, tornou-se possível a criação de regras explícitas ou implícitas (percursos em árvores, por exemplo) permitindo a execução de tarefas de classificação, resumos ou mesmo geração de novos textos. No entanto, a notoriedade do paradigma estatístico se deve primordialmente às aplicações voltadas à tradução automática (Caseli *et al.* 2023).

Em sequência, o contínuo e expressivo ganho de poder de memória e de processamento dos computadores, possibilitou o desenvolvimento (e dominância) da **abordagem conexionista (também chamada de Abordagem emergentista**[7]**. ou paradigma neural)**.

Essa abordagem baseia-se no processamento de grandes quantidades de dados por meio de estruturas (arquiteturas e algoritmos) bastante complexas, como as Redes Neurais Profundas (conhecidas em inglês como *deep learning*).

Da mesma forma que o paradigma estatístico, as redes neurais também se baseiam em grandes volumes de dados para aprender um modelo; contudo, a forma como esse aprendizado é realizado é diferente, uma vez que envolve várias camadas de unidades de processamento para reconhecer os padrões recorrentes (Caseli *et al.* 2023).

Conforme observado por Caseli *et al.* 2023 devido à complexidade das arquiteturas

[7] Informações extraídas do site "O que é inteligência artificial geral?", disponível em https://aws.amazon.com/pt/what-is/artificial-general-intelligence/

compostas por diversas camadas de processamento, a utilização da abordagem conexionista faz com que os fatores associados a um determinado comportamento ou resultado sejam praticamente irrecuperáveis, "tornando o código opaco, e seu efeito, não previsível (não determinístico)". Assim, recentemente, com o objetivo de fornecer alguma explicabilidade aos processos de PLN, vem ganhado espaço abordagens híbridas, que combinam principalmente o paradigma simbólico com uma das outras abordagens comentadas.

2.3.2.1 **Mineração de textos**

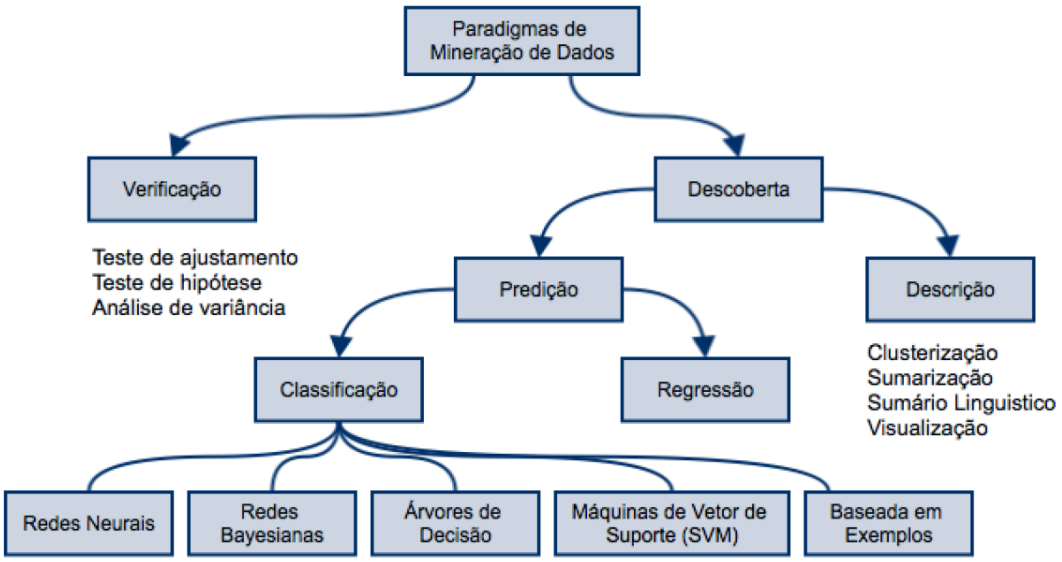
A **mineração de texto (*Text Mining*)**[8], pode ser definida como um processo específico da área de mineração de dados destinado à identificação de padrões potencialmente úteis e compreensíveis embutidos em dados textuais (Hearst 1999; Ebecken *et al.* 2003; Kao e Poteet, 2007; apud Ribeiro 2018).

Vale ressaltar que soluções voltadas à mineração de textos, via de regra, partem de resultados de ferramentas computacionais responsáveis pela transformação de dados textuais brutos (não estruturados ou semiestruturados) em dados estruturados – o que possibilita a posterior utilização de métodos analíticos.

Baseado em proposta de Maimon e Rokach, 2010, os métodos de mineração de textos podem ser divididos conforme os paradigmas representados na Figura a 11.

[8] O conceito de mineração de texto, por vezes, é tomado como sinônimo de análise de texto (*Text Analytics*), entretanto, mais precisamente, pode-se associar a mineração de texto ao processo que combina estatística, linguística e aprendizado de máquina para prever automaticamente resultados de experiências anteriores. Já a análise de texto remete primordialmente a visualizações de dados (como tabelas, gráficos ou outros relatórios visuais) elaboradas a partir dos resultados de análises da mineração de textos, com o intuito de detectar padrões nos dados textuais (Fonte: "What is text mining?", disponível em https://www.ibm.com/topics/text-mining).

Figura 11 – Taxonomia dos métodos de Mineração de Dados



Fonte: Silva apud Maimon e Rokach, 2010

Com relação às tarefas abrangidas pela mineração de texto, pode-se citar: (i) recuperação de informações; (ii) categorização e agrupamento de texto; (iii) reconhecimento de entidades nomeadas; (iv) modelagem de relações entre entidades; (v) produção de taxonomias granulares; (vi) análise de sentimentos; (vii) resumo de documentos; (viii) classificação preditivas de documento; (ix) coleta de registros para bancos de dados; (x) sumarização de conteúdos; e (xi) elaboração de índices de pesquisa, para facilitar o acesso à elementos relevantes de textos.

Dentre as possibilidades da mineração de texto elencadas, a que interessa mais diretamente ao presente trabalho diz respeito à tarefa de recuperação de informações, o que justifica o destaque dado a referida técnica no subtópico a seguir.

2.3.2.2 **Recuperação de informação**

A subárea da mineração de texto denominada recuperação de informação, do inglês *Information Retrieval* (IR), trata da representação, armazenamento, organização e acesso a itens de informação contidas em coleções de documentos (Baeza-Yates 1999).

A recuperação de informação parte de uma necessidade do usuário, tendo por objetivo a apresentação de informações ou documentos relevantes (sejam eles textos, imagens, sons e outros tipos), com base em um **conjunto predefinido de consultas ou frases** – constituindo a essência de soluções que abarcam desde sistemas de catálogo de bibliotecas até mecanismos de busca na web, como o Google.

Neste contexto, para obter documentos de seu interesse, o usuário deverá traduzir uma necessidade de informação em uma consulta. Tal informação (de interesse do usuário)

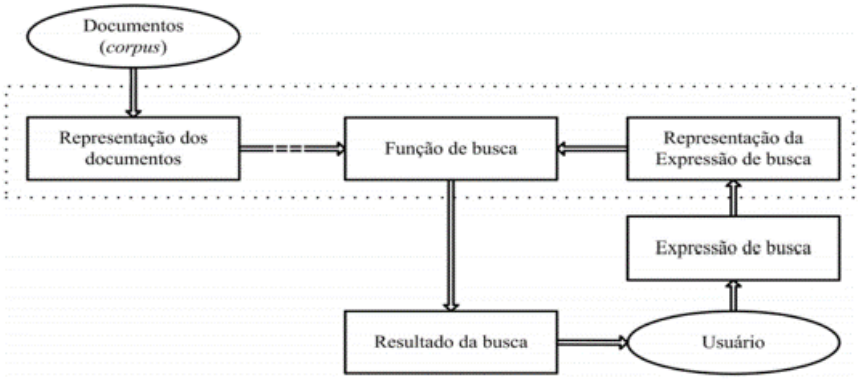
é normalmente chamada de necessidade de informação do usuário.

Para ser eficaz na tarefa de satisfazer a necessidade de informação do usuário, os sistemas de recuperação de informação devem ser capazes de ordenar os elementos de uma coleção de documentos de acordo com o seu grau de **relevância** em relação à consulta elaborada. A noção de relevância, portanto, é um conceito fundamental em recuperação de informação e é um componente chave para calcular a classificação (ordenação) de documentos, com vista a fornecer a resposta ao usuário.

Cabe destacar que, em geral, na área da recuperação da informação não se fale em obter uma resposta "definitivamente correta" a determinada consulta, pois a resposta depende da necessidade de informação do usuário em determinado momento – o que afasta, por exemplo, a técnica em comento da teoria básica de banco de dados e linguagens como SQL (Baeza-Yates 1999 e Salton 1971).

A figura a seguir apresenta o modelo esquemático do processo de recuperação de informação:

Figura 12 – Modelo de recuperação da informação



Fonte: Ferneda 2012

A função de busca de textos, coração do processo ilustrado, tem por objetivo identificar a informação que o usuário necessita, com base na análise do grau de similaridade entre a representação da expressão de busca e as representações dos documentos (a tarefa de representação será objeto do Subtópico 2.3.4, dedicado à semântica distribucional).

Especialmente em razão da necessidade de lidar com o imenso e diversificado volume de documentos disponibilizados pela internet, o desenvolvimento de soluções de representação e recuperação de informações são, atualmente, protagonizadas por técnicas de inteligência artificial.

Um sistema de recuperação de informações pode envolver os distintos níveis linguísticos já comentados neste trabalho (morfológico, léxico, sintático, semântico e pragmático).

Por outro lado, vale anotar que muitas tarefas de processamento de linguagem não requerem tratamentos ou estratégias complexas para o alcance de seus objetivos – podendo ser suficiente, nesses casos, apenas uma análise parcial (ou superficial) de sentenças de entrada. Por exemplo, os sistemas de extração de informação geralmente não extraem todas as informações possíveis de um texto; eles simplesmente identificam e classificam os segmentos de texto que provavelmente contêm informações valiosas. Similarmente, sistemas de recuperação de informações podem optar por indexar ou identificar elementos em documentos com base em um subconjunto selecionado dos constituintes encontrados em um texto, a exemplo de similaridade de sentenças (Jurafsky e Martin, 2008).

2.3.3 **Pre-Processamento**

O pré-processamento de dados abrange operações de preparação, organização e estruturação de dados brutos. Trata-se de uma etapa da mineração de dados – sejam eles numéricos, textuais, categóricos etc. – que normalmente precede a realização de análises e predições propriamente ditas.

Para o presente trabalho interessam especificamente as técnicas aplicáveis na área de análise de dados textuais, assim serão abordados os conceitos básicos de tokenização, remoção de *stop words*, *stemming* e *lemmatization*. Essas técnicas citada são, na maioria dos casos, executadas por meio de funções que constam das bibliotecas em Python especializadas em processamento de linguagem natural, como NLTK (*Natural Language Toolkit*)[9] e SpaCy[10].

2.3.3.1 **Tokenização**

O processo de tokenização (do inglês *tokenization*) tem por objetivo dividir um texto em unidades, que podem ser palavras ou sentenças (que neste trabalho assume-se corresponder a "frases") chamadas "*tokens*".

A fase de tokenização pode estar associada à função de conversão de todos os caracteres do texto de entrada para letra minúscula. A vantagem deste procedimento é fazer com que o tratamento dos tokens não seja influenciado pelo fato de o texto estar escrito em letras maiúsculas ou minúsculas (case-sensitive). No entanto, deve-se ressaltar, que essa opção dificulta a execução de tarefas como Reconhecimento de Entidades Nomeadas, do inglês *Named Entity Recognition* (NER) – uma vez que as entidades a serem reconhecidas como nomes próprios são mais facilmente identificáveis por iniciarem com grafia em letra maiúscula.

A depender dos objetivos almejados, também podem contribuir para a eficiência da tarefa de tokenização a remoção de (i) caracteres de pontuação, e (ii) sinais diacríticos,

[9] Documentação da ferramenta disponível em https://www.nltk.org/
[10] Documentação da ferramenta disponível em https://spacy.io/models/

assim compreendidos os acentos (grave, circunflexo ou til) e sinais como cedilha (Caseli *et al.* 2023).

Toma-se como exemplo a frase, do escritor Terry Pratchett, traduzida livremente para o português: "A estupidez real sempre vence a inteligência artificial". Submetida a um processo de tokenização por palavras, a referida frase de entrada resultaria em oito tokens: "a", "estupidez", "real", "sempre", "vence", "a", "inteligencia", "artificial"[11].

Algumas tarefas de PLN, entretanto, podem ser melhor desempenhadas a partir da tonkenização por sentenças (frases). Sobre esse assunto anota-se que a segmentação de um texto em frases geralmente é baseada na pontuação, visto que certos tipos de caracteres, como os pontos de interrogação e de exclamação, regularmente, são marcadores de limites de frases. O sinal denominado "período", apesar de popularmente chamado de "ponto final" ("."), é mais ambíguo – pois além de demarcarem o limite de uma frase, exerce outras funções no texto, como, por exemplo, marcador de abreviaturas como Sr. ou Ltda. (Jurafsky e Martin, 2008).

2.3.3.2 **Remoção de *stop words***

As *stop words* (sem tradução precisa em português) são palavras que costumam ocorrer com elevada frequência em indistintos tipos de documentos e assim pouco contribuem para apreensão do conteúdo semântico dos textos analisados. Em geral, considera-se stop word: preposições, conjunções, artigos e verbos de ligação.

Para esse processo normalmente são utilizadas funções já implementadas em bibliotecas como o NLTK, que identificam e facilitam o descarte dessas *stop words*, a partir de uma lista de palavras previamente construída – que pode ser adaptada de acordo com os requisitos e objetivos de cada aplicação (Caseli *et al.* 2023).

Conforme Caseli *et al.* 2023, apesar de simples, o processo de remoção de *stop words* exige cuidados e deve considerar o contexto da aplicação desenvolvida, pois pode importar em perda substancial de informações relevantes, conforme exemplifica as autoras:

> Se pensarmos na famosa expressão ser ou não ser, eis a questão", com a remoção de stop words, sobraria apenas questão, o que descaracteriza completamente a expressão. O impacto negativo na busca seria que o sistema não mais conseguiria distinguir entre documentos que contenham a expressão completa daqueles que contêm apenas a palavra "questão".

Como exemplificação da utilização prática da técnica em comento, a já citada frase, "A estupidez real sempre vence a inteligência artificial", submetida a um processo de remoção

[11] Nesse exemplo, destaca-se que o único sinal diacrítico removido foi o acento circunflexo de "inteligência"

de *stop words*, teria como resultado a remoção dos dois artigos "a" pressentes na sua construção original, resultando em : "estupidez real sempre vence inteligência artificial"[12]

2.3.3.3 **Stemming**

O procedimento denominado *stemming* (sem tradução estabelecida em português) se refere basicamente à eliminação dos prefixos e sufixos das palavras para derivar sua forma raiz.

O objetivo do *stemming* é gerar uma mesma representação para formas variantes de uma mesma palavra. Por exemplo, removendo-se os sufixos de "apresentação", "apresentando", e "apresentar", obtém-se o radical "apresent". Assim, distintos termos, derivados de um mesmo radical, passam a ser contabilizados como um único termo (Caseli *et al.* 2023).

Em tarefas de recuperação de informações, essa técnica tende a melhorar os resultados (i) aumentando o número de documentos relevantes recuperados em resposta a uma consulta formulada (Caseli *et al.* 2023); bem como (ii) reduzindo o tamanho dos arquivos de indexação.

Anota-se que para a língua inglesa, o primeiro algoritmo de *stemming* (ou *stemmer*) foi proposto por Julie Beth Lovins em 1968 (Lovins 1968, apud Caseli *et al.* 2023). Em 1980, Martin Porter propôs o *Porter Stemmer* (Porter 1980) que obteve bons resultados para a língua inglesa, sendo, posteriormente, traduzido para outros idiomas, incluindo o português. As autoras Caseli *et al.* 2023, destacam ainda outras iniciativas e resultados de utilização de stemmings em língua portuguesa:

> O primeiro stemmer desenvolvido especialmente para a língua portuguesa foi o Removedor de Sufixos da Língua Portuguesa (RSLP) (Orengo; Huyck, 2001). Posteriormente veio o STEMBR (Alvares; Garcia; Ferraz, 2005). Há resultados experimentais que mostram a validade da aplicação de stemmers para RI em português (Flores; Moreira; Heuser, 2010; Orengo; Buriol; Coelho, 2006). Muitas vezes, na média de um conjunto de consultas, o impacto do stemming pode ser pequeno. Contudo, ao analisarmos consultas individuais, percebemos melhorias muito grandes em alguns casos. Em outros, podem aparecer mais documentos não relevantes recuperados pois a remoção do sufixo pode gerar ambiguidade. Os resultados mostram que stemmers menos agressivos, ou seja, com menos regras de remoção tendem a gerar resultados melhores pois têm menos impacto negativo na precisão.

Aplicando-se o *stemming* na mesma frase que vem sendo adotada como exemplo nos subtópicos anteriores de pré-processamento (itens 2.3.3.2 e 2.3.3.1), tem-se como resultado:

[12] Procedimento realizado por meio de funcionalidade disponível em https://bho.li/tool/stop_words_removal/ (selecionando-se a opção de língua portuguesa)

"A estupid real sempr venc a intelig artific"[13].

2.3.3.4 ***Lemmatization***

A técnica conhecida como lematização (do inglês *lemmatization*) consiste em substituir uma palavra de um texto analisado por sua forma raiz. Por exemplo, as palavras "cantou", "cantado", e "canta" são todas formas do verbo "cantar". Nesse exemplo, o mapeamento das flexões raiz do verbo ("cantar") é chamado de lematização; já a forma raiz obtida a partir deste processo é chamada de *common lemma* (Jurafsky e Martin, 2008).

O objetivo do uso da técnica em comento é melhorar o desempenho de tarefas de PLN, tendo em vista que na maioria das vezes a flexão de gênero, número, tempo não importa significativamente na compreensão do fato/ação objeto de uma sentença.

Deve-se observar que a função de lematização idealmente deve considerar a classe gramatical da palavra analisada para, em seguida, usar essa informação para definir o radical que deve ser utilizado para representá-la – pois o radical adequado para uma palavra geralmente depende de como ela é usada em uma frase.

Aplicando-se a lematização à sentença que já serviu de exemplo nos subtópicos antecedentes, ter-se-ia como resultado: "A estupidez real sempre vencer a inteligência artificial" (trazendo para a forma infinitiva a flexão verbal "vence").

2.3.4 **Semântica Distribucional**

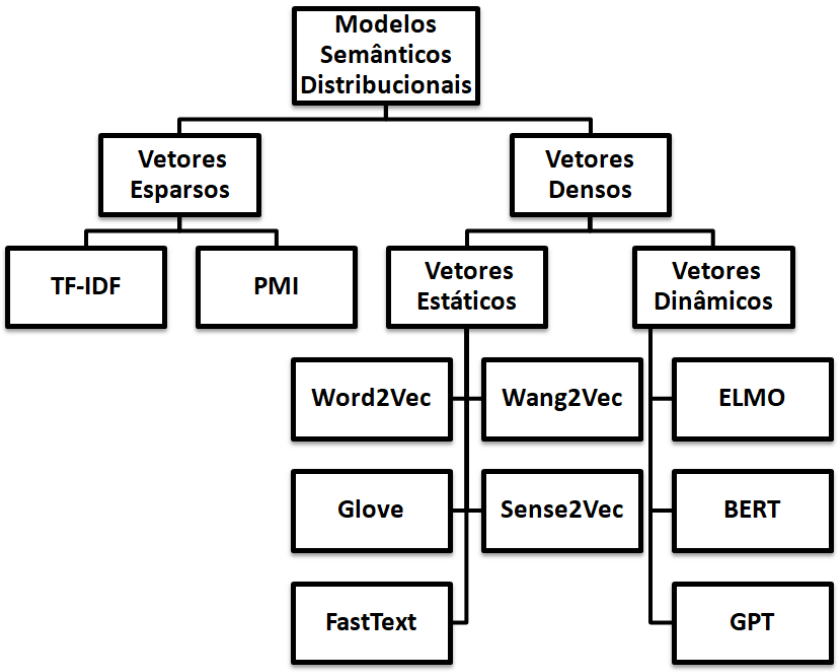
A denominada semântica distribucional é baseada na Hipótese Distribucional (Firth 1935 e Harris 1954) a qual preconiza que palavras em um *contexto linguístico* semelhante tendem a ter significado similar ou aproximado. Assim, de pronto, interessa ressaltar que na semântica distribucional as palavras consideram o contexto em que elas aparecem.

Nessa abordagem, os itens lexicais (palavras) são representados por meio de vetores de valores reais, conhecidos por **vetores semânticos**, que codificam o significado das palavras a partir de sua distribuição nos textos (Caseli *et al.* 2023) – essa estratégia possibilita a realização de tarefas não-supervisionadas (dispensando o processo de rotulagem dos dados).

Os modelos que operam esse tipo de abordagem são chamados de Modelos Semânticos Distribucionais (MSD), *Distributional Semantic Models* (DSM), no inglês, podendo ser classificados, em um primeiro nível, pelo fato de se basearem em vetores esparsos ou vetores densos, conforme ilustra o esquema apresentado por (Caseli *et al.* 2023):

[13] Procedimento realizado por meio de funcionalidade disponível em http://text-processing.com/demo/stem/ (selecionando o stemmer Portuguese RSLP)

Figura 13 – Ilustração dos modelos semânticos distribucionais



Fonte: Caseli *et al.* 2023

No caso dos vetores esparsos (assim entendidos os vetores compostos por grande quantidade de elementos com valor zero) o significado de uma palavra é representado com base na frequência da palavra em uma coleção de documentos.

Um exemplo de modelo que se baseia em vetores esparsos é o *Term Frequency – Inverse Document Frequency* (TF-IDF), utilizado em tarefas que envolvem a comparação de similaridade entre documentos, como a detecção de plágio, a inferência textual e a recuperação de informações (Caseli *et al.* 2023).

Nos vetores densos, a representação da informação é codificada por números reais em menores dimensões vetoriais, já que possuem poucas posições zeradas. Na área de processamento de linguagem natural, vetores densos correspondem ao conceito de *embeddings*, ao qual é dedicado o subtópico seguinte.

2.3.4.1 ***Embeddings***

embeddings (termo que poderia ser aproximado para o português como "incorporação de palavras") são vetores densos e de reduzida dimensão, cuja notória vantagem é o aumento da capacidade computacional, quando comparada ao processamento para se manipular vetores dispersos e de alta dimensionalidade (Goldberg 2022).

A grande evolução atribuída à utilização dos *embeddings* diz respeito à capacidade de captura das relações semânticas e contextuais entre as palavras – conforme já apresentado na Figura 13, esses modelos podem ser subdivididos em vetores estáticos e dinâmicos.

As representações baseadas em **vetores estáticos (*Static Word Embeddings*)**, funcionam como um mapeamento de palavras para vetores, de modo que o modelo em si, após o treinamento para obtenção dos vetores, é descartado. Assim, os *embeddings* estáticos permanecem fixos uma vez aprendidos, ou seja, eles não podem ser ajustados ou modificados para uma tarefa específica.

O contexto, no qual uma determinada palavra está inserida, é considerado apenas para o treinamento do modelo. Assim, na utilização do modelo (pré-treinado), para converter as palavras de um documento em vetores, as palavras vizinhas ou contexto do referido documento não são computadas, impossibilitando a obtenção, por exemplo, de diferentes vetores para palavras homônimas.

Já os ***embeddings* dinâmicos**, também conhecidos como *embeddings* contextuais (em inglês, *Contextual Embeddings*), podem se adaptar às nuances específicas de uma tarefa e ao contexto em que as palavras estão inseridas (Caseli *et al.* 2023). Dentre os modelos baseados em vetores densos dinâmicos cita-se o *Bidirectional Encoder Representations from Transformers* (BERT), ao qual foi dedicado tópico específico deste trabalho.

Neste caso, a representação vetorial considera o contexto do texto, obtendo diferentes vetores para uma mesma palavra para cada vez em que ela é acompanhada por um distinto conjunto de palavras, de modo a representar seus diferentes significados (Jurafsky e Martin, 2008).

Tais representações conseguem capturar várias propriedades sintáticas e semânticas, de modo que elas atingem performance no estado da arte em uma grande variedade de tarefas de PLN (Liu *et al.* 2020). No entanto, como os modelos precisam observar todas as palavras da sentença, essas representações envolvem alto custo computacional.

Importa ressaltar que as representações contextuais dependem da utilização de um **modelo de linguagem** treinado para calcular, no momento de análise, o vetor referente à palavra (Sieg 2019). Um modelo de linguagem pode ser compreendido como uma distribuição de probabilidades que prediz a possibilidade de determinada sequência de palavras aparecer na linguagem de estudo, sendo obtido por métodos não supervisionados de aprendizagem de máquina (Liu *et al.* 2020).

Modelos como ELMo (Peters *et al.* 2018), GPT (Radford *et al.* 2018) e BERT (Devlin *et al.* 2019) são capazes de gerar representações vetoriais contextuais utilizando um **modelo de linguagem neural** (Gomez-Perez *et al.* 2020).

2.3.4.2 Busca Semântica

A denominada busca semântica, vai além da busca tradicional baseada em palavras-chave, permitindo que os usuários encontrem informações a partir dos significados dos termos consultados (ainda que não haja exata correspondência com palavras da fonte de

dados pesquisada), fornecendo, assim, uma gama mais ampla de possibilidades na execução de tarefas em PLN.

Na geração de textos e na tradução automática, por exemplo, saber se duas palavras são semelhantes semanticamente, em um determinado contexto, torna-se essencial para decidir se é possível substituir uma pela outra. Também pode-se usar similaridade semântica, por exemplo, para realizar agrupamento de palavras ou para graduar as respostas de estudantes em um exame de redação (Jurafsky e Martin, 2008).

As tarefas de recuperação de informações ou resposta a perguntas formuladas também requerem a identificação de documentos cujas palavras tenham significados (semântica) semelhantes aos termos de consulta. Nesse sentido, frise-se que diversos trabalhos abordam as vantagens de usar informação semântica para melhorar o desempenho de sistemas de recuperação de informação (Han e Srihari, 1999, Voorhees 1993, Arampatzis *et al.* 2000, apud Loper 2000).

2.3.4.3 **Similaridade de cosseno**

A capacidade de calcular a similaridade semântica de palavras, sentenças ou textos é um dos recursos mais relevantes do processamento de linguagem natural. Para tanto, os conteúdos na forma textual precisam primeiro ser convertidos para suas representações numéricas vetoriais (*embedding*) por meio da utilização de um modelo de linguagem contextual (a exemplo do BERT comentado neste capítulo). Tais vetores servem como entrada para uma função de busca por elementos próximos no espaço vetorial.

A similaridade de cosseno (do inglês *Cosine-Similarity*) é uma medida de distância entre dois vetores de um dado espaço vetorial, baseada no valor do cosseno do ângulo compreendido entre eles. Sua fórmula corresponde a:

$$scos(\vec{f}, \vec{v}) = \frac{\vec{f} \cdot \vec{v}}{\left|\vec{f}\right| |\vec{v}|} = \frac{\sum_{i=1}^{n} f_i v_i}{\sqrt{\sum_{i=1}^{n} f_i^2} \sqrt{\sum_{i=1}^{n} v_i^2}}$$

A similaridade por cosseno é usada em várias soluções em que a magnitude dos vetores não é tão importante quanto sua direção. O cálculo apresentado é baseado na operação de produto escalar, da álgebra linear. Em suma, resultados com valores próximos de 1 representam vetores que apontam para direções semelhantes, o que indicia, para solução desenvolvida neste trabalho, sentenças similares.

Vale o registro que as bibliotecas mais utilizadas para PLN, a exemplo da NLTK (já comentada anteriormente), possuem funções pré-definidas que retornam os valores de similaridade de forma prática e eficiente.

3 METODOLOGIA

Tendo em vista o objetivo deste trabalho, que em essência se constituiu na implementação de uma ferramenta de recuperação de informações originárias de deliberações do TCU, a primeira etapa de desenvolvimento enfrentada referiu-se à obtenção das bases de dados textuais que servem como o repositório destas deliberações[1].

O Tribunal de Contas da União disponibiliza arquivos de dados textuais relativos às deliberações publicadas por ano pelo órgão (desde o ano de 1992). O acesso a tais recursos é realizado por meio de downloads, para os quais não é necessário acesso privilegiado ou cadastro prévio, em página web específica para este fim[2].

Os arquivos disponibilizados pelo TCU têm formato ".csv" delimitados pelo caractere "|" (ASCII 124), contendo ao todo 33 colunas[3]. O órgão disponibiliza ainda um dicionário dos metadados dos arquivos em comento[4].

Para manipulação dos arquivos ".csv" optou-se por sua leitura e conversão para o formato *Dataframe* do Pandas[5]. A utilização da referida biblioteca deu-se por sua ampla difusão, pela disponibilidade de métodos que serviram para execução das distintas etapas pré-processamento, bem como pelas facilidades de integração com outras bibliotecas.

3.1 Seleção das informações de interesse

Para o presente trabalho foram selecionados apenas os bancos de dados que contemplam as deliberações relativas aos anos de 2023 e 2024, por representaram posicionamentos jurisprudenciais mais atualizados do TCU e versarem sobre legislação mais recente – a exem-

[1] Importa esclarecer que o documento que neste texto está sendo referenciado pelo termo "deliberação", mais comumente é conhecido pelo termo "acórdão". *Stricto sensu*, entretanto, "acórdão" corresponde a uma das partes constituintes da deliberação (sendo as outras partes o "relatório" e o "voto"). Opta-se por essa terminologia mais rigorosa (sentido específico) para evitar imprecisões nas referências feitas ao conteúdo desses elementos no decorrer do texto.

[2] Disponível em https://sites.tcu.gov.br/jurisprudencia/index.html, opção "Acórdãos – por ano" (reforça-se que na terminologia adotada neste trabalho tais documentos correspondem à "deliberações")

[3] As denominações das colunas dos arquivos disponibilizados são: KEY; TIPO; TITULO; NUMACORDAO; ANOACORDAO; NUMATA; COLEGIADO; DATASESSAO; RELATOR; SITUACAO; PROC; ACORDAOSRELACIONADOS; TIPOPROCESSO; INTERESSADOS; ENTIDADE; RELATORDELIBERACAORECORRIDA; MINISTROREVISOR; MINISTROAUTORVOTOVENCEDOR; REPRESENTANTEMP; UNIDADETECNICA; ADVOGADO; ASSUNTO; SUMARIO; ACORDAO; DECISAO; QUORUM; MINISTROALEGOUIMPEDIMENTOSESSAO; RECURSOS; RELATORIO; VOTO; DECLARACAOVOTO; VOTOCOMPLEMENTAR; VOTOMINISTROREVISOR

[4] Disponível em https://sites.tcu.gov.br/jurisprudencia/dicionario-dados.html

[5] Disponível em https://pandas.pydata.org/

plo dos novos arcabouços legais que passaram a reger as licitações públicas em substituição à Lei 8.666/1993 (que perdeu sua vigência em definitivo a partir de 30/12/2023[6]).

Em sequência, a partir de uma análise exploratória inicial, foi verificado que apenas algumas das colunas das bases de dados selecionadas seriam, de fato, relevantes aos objetivos deste trabalho, são essas apresentadas na Tabela 1.

Tabela 1 – Informações sobre as colunas do dicionário de dados

Nome da coluna (chave do dicionário)	Tipo do dado	Descrição do conteúdo[1]	Exemplo obtido da base real (valores do dicionário)
KEY	String que segue padrão baseada em elementos de identificação do acórdão	Identificador único para a deliberação.	"ACORDAO-COMPLETO-2623236"
UNIDADETECNICA	String que segue padrão baseada nos nomes das unidades técnicas à época da deliberação	Unidade do TCU responsável pela análise.	"Unidade de Auditoria Especializada em Infraestrutura Portuária e Ferroviária (AudPortoFerrovia)"
TIPOPROCESSO	String que segue baseada nos tipos de processos	Tipo do processo que originou o Acórdão.	"RELATÓRIO DE AUDITORIA (RA)"
RELATORIO	Texto longo com marcações HTML	Descreve as informações preliminares do caso que está sendo analisado.	<p class=\"TCU_-_-Epígrafe\">\tAdoto como Relatório transcrição do relatório (...)"!
VOTO	Texto longo com marcações HTML	Voto do relator ou de outros ministros presentes na sessão.	"<p class=\"TCU_-_-Epígrafe\">Cuidam os autos de relatório de auditoria realizada na (...)"
ACORDAO	Texto longo com marcações HTML	Texto completo e detalhado do acórdão.	"<p class=\"TCU_--_Ac_-_item_9_--_§§\">VISTOS, relatados e discutidos estes autos de auditoria (...)"

[1]Descrição obtida do dicionário de dados dos bancos de dados fornecido no site na página do TCU, disponível em: https://sites.tcu.gov.br/jurisprudencia/dicionario-dados.html
Fonte: elaboração própria

Conforme registrado na introdução deste trabalho, o TCU tem uma ampla gama de tipo de processos, incluindo aqueles cujo objeto envolve assuntos muito específicos.

[6] O artigo 191 combinado com o artigo 193 da Lei 14.133/2021 estabeleceu regime de transição do regramento de licitações anterior para o atual, fixando a data 29/12/2023 como o último dia de vigência das Leis 8.666/93, 10.520/2002 e 12.462/2011. Atualmente, além da própria Lei 14.133/2021, as licitação públicas também são regidas pela Lei 13.303/2016, aplicável a empresas públicas estatais.

Pontua-se, por exemplo, que numerosos processos do TCU dizem respeito à verificação da regularidade de atos de admissão e de aposentadoria de servidores de toda a administração pública federal – sendo a análise deste tipo de tema bastante especializada e baseada em referências que, a princípio, não seriam úteis aos objetivos pretendidos neste trabalho.

Nessa esteira, para seleção das deliberações que compõem os *datasets*, sobre os quais opera a ferramenta de recuperação de informações tratada neste trabalho, foram utilizados dois critérios atinentes: (i) ao tipo do processo da deliberação; e (ii) a unidade técnica que atou no processo.

Quanto à **seleção relacionada ao tipo do processo**, o objetivo foi que as deliberações abrangidas pelo escopo do trabalho dissessem respeito às áreas finalísticas do TCU especializadas em temas **relacionados à licitação e à contratação de obras públicas de infraestrutura**.

Assim, foi realizada **seleção baseada nas unidades técnicas** (que atuam na fase inicial dos processos, anteriormente ao envio desses ao gabinete do ministro relator), optando-se por selecionar aquelas cujas competências dizem respeito diretamente à fiscalização de obras na área de infraestrutura. Para tanto, foi aplicado filtro à coluna "UNIDADETECNICA" dos *datasets* baseado nas referências apresentadas na Tabela 3.

Tabela 2 – Unidades técnicas selecionadas e expressões utilizadas para identificá-las no respectivo campo do dataset

Denominação atual da unidade selecionada[1]	Denominação anterior da unidade selecionada[2]	Expressão (radical) utilizadas para seleção
Unidade de Auditoria Especializada em Infraestrutura Urbana e Hídrica (SeinfraUrb)	Secretaria de Fiscalização de Infraestrutura Urbana (SeinfraUrbana)	"SeinfraUrb" e "AudUrb"
Unidade de Auditoria Especializada em Infraestrutura Rodoviária e Aviação Civil (AudRodoviaAviação)	Secretaria de Fiscalização de Infraestrutura Rodoviária e de Aviação Civil (SeinfraRodoviaAviação)	"SeinfraRod" e "AudRod"
Unidade de Auditoria Especializada em Infraestrutura Portuária e Ferroviária (AudPortoFerrovia)	Secretaria de Fiscalização de Infraestrutura Portuária e Ferroviária (SeinfraPortoFerrovia)	"SeinfraPor" e "AudPor"

[1]A estrutura do TCU e as competências das unidades estão definidas na Resolução-TCU 347, de 12/12/2022.
[2]Uma explicação detalhada que relacionada as denominações da estrutura anterior e atual é apresentada em página específica do TCU, disponível em: https://portal.tcu.gov.br/imprensa/noticias/conheca-a-nova-estrutura-da-secretaria-geral-de-controle-externo-area-fim-do-tcu.htm
Fonte: elaboração própria

Ainda com respeito aos critérios de seleção de deliberações baseada no tipo do processo, esclarece-se sobre existência de processos, que, apesar de dizerem respeito às áreas finalísticas do TCU, não possuem informações relativas aos campos "relatório" e

"voto" (como é o caso de processos classificados como "Representação" ou "Denúncia", por exemplo). Com base nestas considerações, optou-se por selecionar apenas **processos do tipo "Relatório de Fiscalização"**, implementando-se a seleção a partir de filtro baseada no radical "RELATÓRIO" no campo "TIPOPROCESSO" das bases de dados.

Conforme já comentado, as etapas de seleção de colunas e de instâncias de interesses se deram com o uso da biblioteca Pandas. Acresce-se que para cada etapa realizada, foram utilizados métodos da citada biblioteca para salvar em disco os resultados obtidos em formato ".json". Também foi elaborado código que possibilita que futuros processamentos sejam realizados a partir da leitura destes arquivos – o que permite executar o modelo partir das etapas intermediárias já processadas em sessões anteriores.

3.2 Extração de informações semiestruturada da base de dados

Superada a etapa de seleção das informações de interesse, foram realizados os procedimentos de pré-processamento dos textos de longa extensão contidos nos campos "RELATORIO", "VOTO" e "ACORDAO". As informações textuais dos citados campos são originalmente formatadas em Linguagem de Marcação de HiperTexto (HTML), conforme se pode observar nos exemplos trazidos na última coluna da Tabela 1.

Assim, o pré-processamento dos campos em comento foi iniciado pela **extração dos textos inseridos nos elementos HTML**, por exemplo, marcações de abertura ("<p>") e fechamento ("<\p>") de parágrafos. A extração do conteúdo textual do formato HTML é realizada por meio do uso da biblioteca BeautifulSoup[7](especializada em extração de dados de arquivos HTML e XML). A saída da função de extração do conteúdo textual é uma variável do tipo lista (*list*), cujos elementos são sentenças.

A etapa seguinte foi dedicada ao **pré-processamento** da saída comentada no parágrafo anterior, subdividindo-se em: (i) remoção de caracteres especiais; (ii) tokenização; (iii) remoção de stop words; e (iv) steemização.

As finalidades desses procedimentos de pré-processamento foram explicadas na fundamentação teórica do presente trabalho, cabendo neste momento detalhar as especificações dos métodos aplicados – implementadas com a utilização do *framework Natural Language Toolkit*[8], considerando:

(a) a remoção de caracteres especiais alcança apenas a expressão "\t", que consta ao início de cada parágrafo do texto original;

(b) para a tokenização foi adotada a classe do NLTK "word_tokenize" que realiza o referido procedimento por palavra do texto (o que resultou em uma lista de tokens

[7] Disponível em https://pypi.org/project/beautifulsoup4/
[8] Disponível em https://www.nltk.org/

que serviu como entrada dos métodos apresentados nas alíneas em sequência);

(c) a remoção de *stop words* foi processada a partir da pesquisa e exclusão de palavras elencadas como *stop words* pela própria biblioteca NTLK[9]; e

(d) para a transformação *steem* (também chamada de radicalização) foi utilizada a classe do NLTK[10] que realiza a tarefa pretendida para a língua portuguesa.

Importa ressaltar que a etapa de pré-processamento contemplou também a reformatação dos dados em uma nova estrutura que serviu de base para todas as etapas posteriores, conforme se descreve a seguir:

(a) as colunas "RELATORIO", "VOTO" e "ACORDAO" (cujo conteúdo foi apresentado na Tabela 1), serviram para formação das respectivas colunas intituladas "RELATORIO_HTML", "VOTO_HTML" e "ACORDAO_HTML" – que têm formato de dicionários cujos elementos foram as sentenças do texto original (com marcações HTML), no seguinte padrão: "0": "sentença a", "1": "sentença b", (...); e

(b) as sentenças pré-processadas, associadas a cada elemento citado na alínea anterior, foram armazenadas nas novas colunas do *Pandas Dataframe* "RELATORIO_TEXTO", "VOTO_TEXTO" e "ACORDAO_TEXTO", que também têm formato de dicionários no seguinte padrão: "0": "sentença pré-processada a", "1": "sentença pré-processada b", (...).

Na forma como foram estruturados os dados, os elementos descritos na alínea "b" anterior serviram, de fato, para a ferramenta de recuperação de informações desenvolvida neste trabalho. No entanto, tais informações possuem pouca inteligibilidade para o usuário (em razão de terem sido submetidos a pré-processamento). Assim, a estrutura de dados adotada possibilita ao usuário acesso aos parágrafos do texto original (descritos na alínea "a"), por meio da equivalência das chaves dos dicionários das colunas apresentadas (cujas denominações incluíram os termos "HTML" e "TEXTO").

3.3 Formação dos repositórios de texto específicos

Os textos das deliberações sobre os quais opera o algoritmo de recuperação de informações para auxílio à escrita foram separados em três *datasets* (que neste trabalho são frequentemente referenciados pelo termo "repositórios").

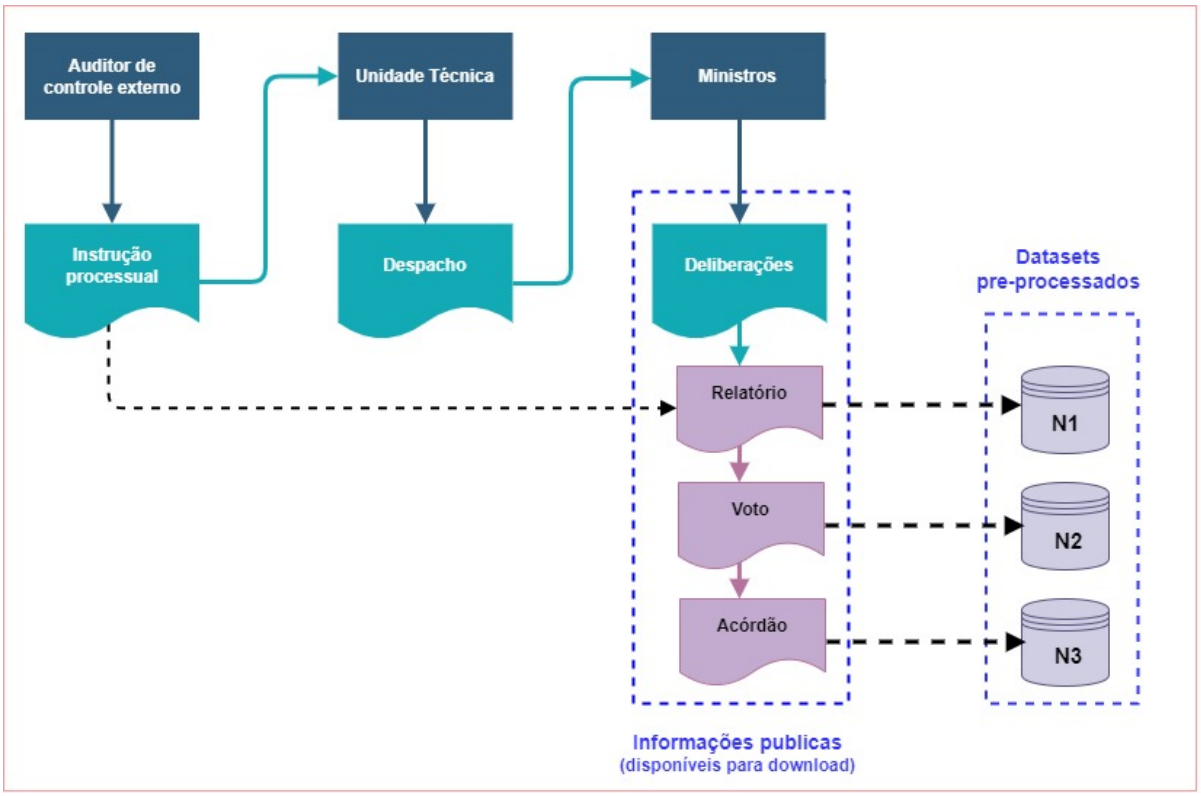
A formação dos repositórios considerou as três partes constituintes das deliberações do TCU, cujos conteúdos estão relacionados às competências das instâncias do órgão

[9] Sintaxe: "nltk.corpus.stopwords.words('portuguese')"

[10] Sintaxe: nltk.stem.RSLPStemmer()

envolvidas no processo de análise e julgamento, conforme ilustrado na Figura 14 com esclarecimentos em sequência.

Figura 14 – Esquema ilustrativo do processo de funcionamento da atuação padrão do TCU e formação dos repositórios



Fonte: Elaboração própria

O **primeiro nível (denominado N1)** é constituído a partir dos dados referentes ao relatório da deliberação. Deve-se esclarecer que a elaboração desse relatório constitui competência do gabinete do ministro relator do processo. No entanto, frequentemente o referido documento é formulado a partir de reproduções de trechos (ou da totalidade) da peça processual de análise técnica elaborada no âmbito das Unidades Técnicas do TCU.

Assim, na maioria das vezes, o *dataset* em comento vem a ser composto de elementos textuais elaborados no âmbito das unidades técnicas, servindo à recuperação de informações estreitamente associadas às recorrentes análises a cargo destas unidades. Noutras palavras, o conteúdo do Repositório N1 tende a estar mais associado ao repertório, ao estilo de discurso e às abordagens que os integrantes do corpo técnico do TCU (Auditores Federais de Controle Externo, lotados nas Unidades Especializadas) registram nas suas análises.

O **segundo nível (denominado N2)** diz respeito ao voto, parte da deliberação, elaborada no âmbito dos gabinetes do ministro relator do processo. Esses documentos tendem a ter linguagem mais sucinta e a abordar de forma mais objetiva e direta apenas

os principais aspectos tratados no processo sob análise (referenciando, quando necessário, os esclarecimentos que constam do relatório).

Esclarece que o voto em comento tem por objetivo apresentar aos demais ministros do TCU (nessa situação normalmente designados pelo termo "pares") as considerações do ministro relator do processo[11] sobre os aspectos mais relevantes trazidos no corpo do relatório. O voto, portanto, ainda não constitui a decisão final sobre o assunto avaliado – o que ainda depende do convencimento da maioria do colegiado composto por nove ministros.

Desta forma, o conteúdo do voto tende a fornecer um posicionamento ou uma análise mais coesa a respeito de determinada questão, o que pode ser de grande valia para o usuário da ferramenta de recuperação de informação objeto deste trabalho.

O **terceiro nível (denominado N3)** trata de posicionamento final do colegiado do TCU, com base no relatório e no voto apresentados pelo ministro relator. Esses acórdãos representam elementos de grande relevância para atuação técnica do TCU, pois registram a palavra final sobre determinado aspecto avaliado, podendo, por exemplo, complementar ou contrariar posicionamento das unidades técnicas – constituindo importantes precedentes sobre determinado assunto que devem ser considerados em futuras análises.

Conforme as premissas apresentadas, os repositórios N1, N2 e N3 foram salvos em arquivos formato ".json" que serviram de insumo às próximas etapas do modelo, contendo as colunas da Tabela 3.

Tabela 3 – Campos dos repositórios utilizados

DB	KEY	UNIDADETECNICA	TIPOPROCESSO	RELATORIO_HTML	RELATORIO_TEXTO	VOTO_HTML	VOTO_TEXTO	ACORDAO_HTML	ACORDAO_TEXTO
N1	x	x	x	x	x				
N2	x	x	x			x	x		
N3	x	x	x					x	x

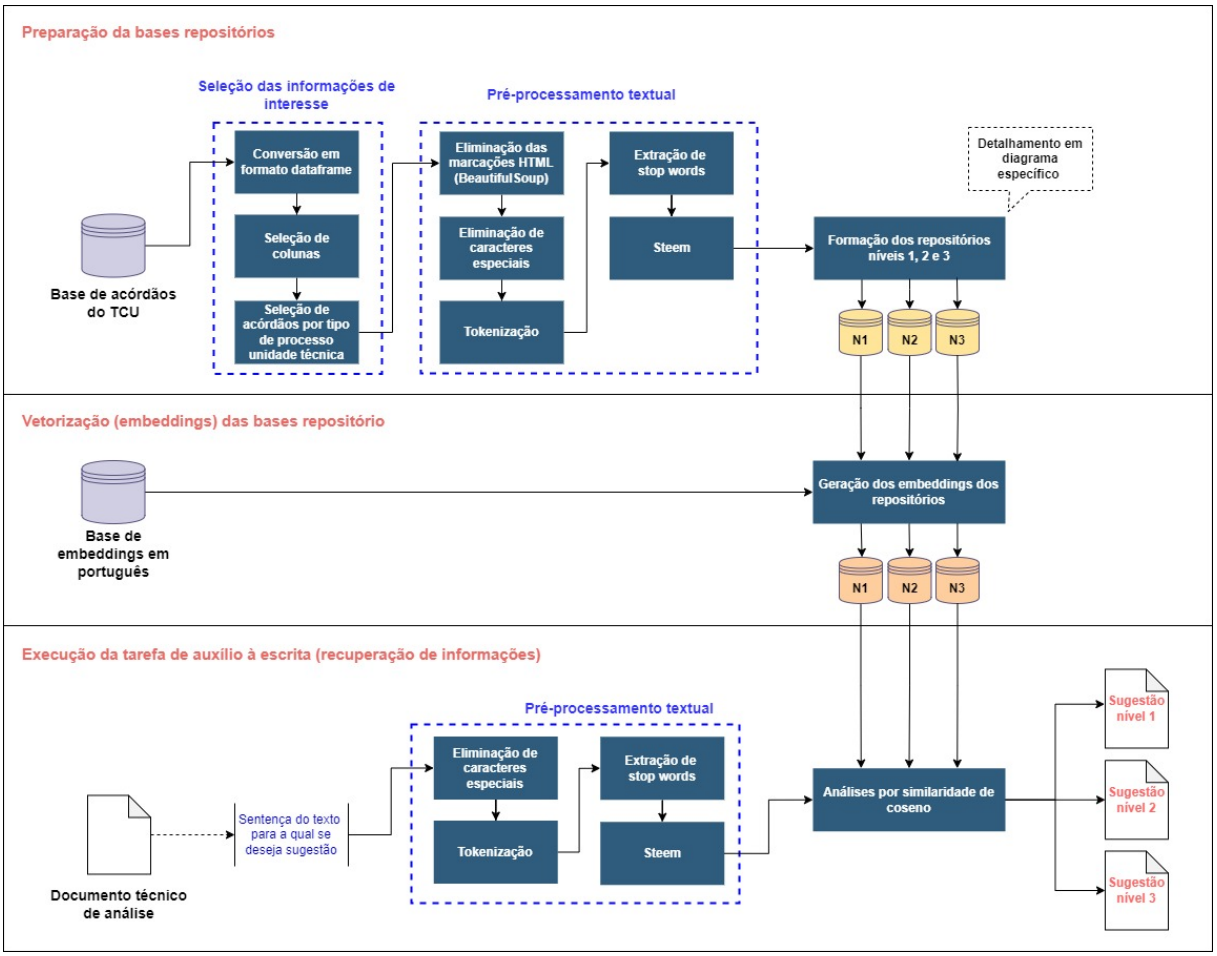
Fonte: elaboração própria

3.4 Implementação do modelo de recuperação de informação

Após as etapas de pré-processamento apresentadas, os próximos passos do desenvolvimento do trabalho tiveram por objeto a implementação do modelo de recuperação de informações propriamente dito, ilustrado na Figura 15 e comentado em sequência.

[11] A partir de 2023, o Tribunal de Contas da União alterou a sistemática de distribuição de processos aos ministros e ministros-substitutos (o que define a relatoria dos processos) e passou a adotar o sorteio como mecanismo primordial de definição de relatoria, com base no disposto na Resolução-TCU 346.

Figura 15 – Esquema geral da ferramenta para recuperação de informações de deliberações do TCU



Fonte: Elaboração própria

Registra-se que o encode dos textos em vetores *embeddings* foi realizado com o uso do *framework* do *Python*, *SentenceTransformers* (baseado na biblioteca *PyTorch*) – que implementa o método descrito no *paper "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks"*, tratado na fundamentação teórica deste trabalho.

Como modelo pré-treinado de *embeddings* em língua portuguesa, compatível com o *SentenceTransformers*, foi adotado o "bert-base-portuguese-cased-nli-assin-2", que conforme documentação técnica

4 AVALIAÇÃO EXPERIMENTAL

O pleno atingimento dos objetivos da presente pesquisa envolve a avaliação da eficiência da solução de recuperação de informações descrita no capítulo dedicado à metodologia. Neste caso, a eficiência avaliada está vinculada à identificação de referências textuais que possam contribuir para elaboração de análises técnicas inseridas nas competências finalísticas do TCU.

Ocorre que, dada a inexistência de uma base de dados de sentenças pré-classificadas (anotadas) com base nos seus respectivos assuntos, que pudesse servir de parâmetro para confrontar com os resultados obtidos, tornou-se impossível a utilização das métricas tradicionalmente utilizadas na validação de resultados de outras tarefas de IA (tais como exatidão, precisão, revocação etc.).

Tampouco foi identificada na literatura acadêmica trabalhos, de escopo e objetivos correlatos à presente pesquisa, dos quais se pudesse aproveitar da metodologia de avaliação para validar os resultados obtidos com a consistência e objetividade desejada.

Neste contexto, a própria a avaliação dos resultados da solução desenvolvida se impôs como um dos relevantes desafios desta pesquisa. Fez-se necessário assim adotar um método específico que – a despeito de não ter a propriedade de "comprovar" – pudesse "indicar" a performance da solução implementada a partir de testes e apontamentos sobre os resultados obtidos.

Desta forma, foram realizadas simulações com o algoritmo de busca de similaridade, utilizando como entrada sentenças de um dos próprios repositórios utilizados na presente pesquisa (denominados N1, N2 e N3, conforme anotado na metodologia), sendo a similaridade avaliada com base em outro dos repositórios citados. Essas simulações possibilitaram duas abordagens de avaliação, descritas em sequência.

4.1 Indicativos de eficiência baseados na convergência entre as deliberações das sentenças de entrada e de saída

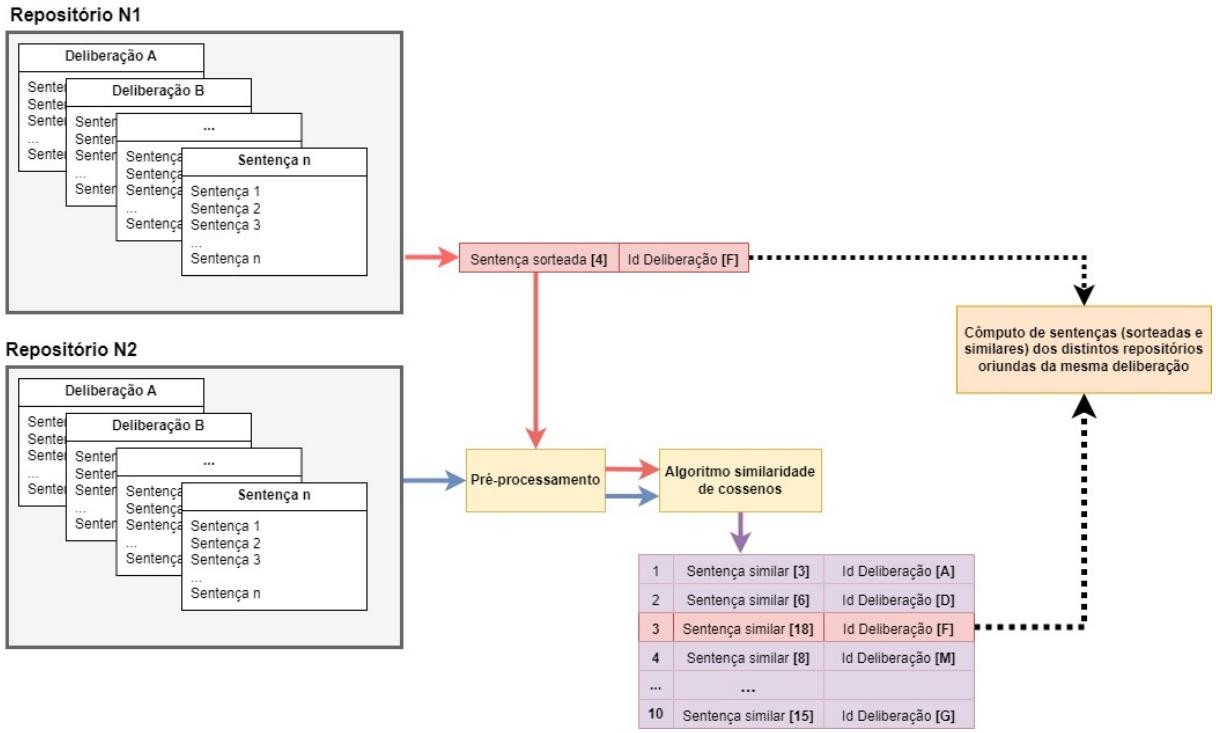
Relembre-se que os *datasets* repositórios dos textos de referência foram elaborados a partir das distintas partes textuais que compõe as deliberações do TCU, conforme registrado na metodologia. Após os pré-processamentos realizados esses repositórios armazenam as seguintes quantidades de sentenças: (i) N1 (textos dos relatórios das deliberações), 30.676 sentenças; (ii) N2 (textos dos votos das deliberações), 4.752 sentenças; e (iii) N3 (acórdãos das deliberações), 1.184 sentenças.

A primeira avaliação indicativa de eficiência da solução implementada fez uso de sentenças oriundas de uma das partes de determinada deliberação (por exemplo, uma

sentença de um relatório armazenada em N1) para servir de entrada para o algoritmo de busca de similaridade em um distinto repositório (por exemplo N2), com o fim de computar em quantas simulações (cujos parâmetros são detalhados mais adiante) as sentenças de entrada e de saída pertenciam à mesma deliberação.

Dada a necessidade de se deixar assente essa metodologia, elaborou-se a Figura 16 que ilustra o procedimento descrito:

Figura 16 – Esquema ilustrativo do procedimento para avaliação de resultados com base na convergência entre as deliberações das sentenças de entrada e de saída



Fonte: Elaboração própria

Partiu-se da premissa que encontrar similaridades entre sentenças de distintas *databases*, porém oriundas da mesma deliberação (que tem um escopo de questões bem delimitado), indiciaria que o algoritmo identificou precisamente o mesmo assunto ou contexto latente da sentença fornecida como entrada – o que, ao menos em tese, serviria como parâmetro que corroboraria o comportamento desejado da solução.

No entanto, deve-se ressalvar que a não inclusão de resultados advindos de uma mesma deliberação da sentença de entrada não aponta necessariamente para imprecisão da solução avaliada. Isto porque é comum, por exemplo, que trechos de uma questão acessória tratada no âmbito de um relatório não ganhe destaque na discussão formulada no voto da própria deliberação. No entanto, é possível que esse mesmo assunto adquira grande importância para construção de um voto associado a um outro processo julgado. Essa compreensão é fundamental para leitura dos resultados apresentados mais adiante.

Baseados nos esclarecimentos apresentados, para avaliação dos resultados foram selecionadas cem sentenças de cada um dos repositórios (parâmetro determinado sem embasamento estatístico) que serviram como entrada para o algoritmo de busca de similaridade descrito no capítulo dedicado à metodologia deste trabalho.

Também de forma empírica, restringiu-se a saída do algoritmo (sentenças similares identificadas) com base em dois critérios: (i) quantidade máxima de dez resultados (ii) similaridade de cossenos igual ou superior a 0,7. Aplicados esses critérios de forma cumulativa, em alguns casos a quantidade de resultados, restrita a dez, impediu que fossem retornadas todas as sentenças com similaridades superiores a 0,7. Noutros casos, o limiar de similaridade fez com que fossem retornados como saída menos de dez resultados.

Desta forma, procedeu-se a seis execuções do algoritmo (simulações) entre cada uma das cem sentenças sorteadas, oriundas de um dos repositórios, com o conjunto de sentenças dos demais repositórios: N1xN2; N1xN3; N2xN1; N2xN3; N3xN1; e N3xN2 (onde lê-se: Sentença de X versus Repositório Y).

Tendo em vista que a cada execução do algoritmo os resultados variaram um pouco (visto que as cem sentenças de entrada são selecionadas por sorteio, com repetição), repetiu-se todas as simulações cinco vezes e considerou-se para fins de resultados a média dessas rodadas.

A tabela 4 apresenta o número de vezes em que dentre os resultados, obtidos por similaridade, foi retornada ao menos uma sentença similar, armazenada em distinto repositório, porém oriunda da mesma deliberação:

Tabela 4 – Porcentagem de sentenças cuja busca de similaridades em repositório resultaram em outras sentenças da mesma deliberação

		Dataset de pesquisa(repositório)		
		N1	N2	N3
Dataset da sentença de entrada	N1	X	47.8%	24.8%
	N2	68.4%	X	27.6%
	N3	55.6%	37.2%	X

Fonte: elaboração própria

Para esclarecer os resultados da tabela 4, toma-se como exemplo o cruzamento da linha N1 com a coluna N2, que indica que para 47,8% das sentenças de entrada, retiradas do tópico "relatório" de uma dada deliberação (repositório N1), alcançaram resposta de similaridade com, ao menos, uma sentença da parte "voto" da mesma deliberação (armazenadas em N2).

Conforme já comentado, não foi identificada uma metodologia sedimentada que permita avaliar de forma objetiva os resultados apresentados na tabela 4. No entanto, pode-se considerar a média global dos experimentos narrados, correspondente a 43,56%, é

um parametro bastante expressivo – dada a baixa possibilidade dessa porcentagem ser alcançada de forma aleatória, em face das dimensões dos repositórios apresentadas ao início deste subtópico.

Frise-se que não seria esperado (e tampouco desejado) que os percentuais em comento se aproximassem de 100%, pois se estaria assumindo um modelo que se mostraria ineficaz para prever resultados inovadores – poder-se-ia aqui fazer uma analogia desta hipótese com o *overfitting* (sobreajuste), tratado no campo das experimentações baseadas em *machine learning*.

Do exposto, pode-se afirmar que os resultados obtidos pelas simulações apresentadas no presente subtópico denotam que o tema das sentenças comparadas entre os repositórios está sendo apreendido pela solução de busca. Considera-se, portanto, que os resultados obtidos podem ser lidos de forma promissora, especialmente, ao serem associados à avaliação apresentada no subtópico seguinte.

4.2 Indicativos de eficiência baseados na análise subjetiva de resultados

Devido às limitações da interpretação dos resultados apresentados no subtópico anterior, entendeu-se pertinente realizar complementarmente uma avaliação subjetiva (executada pelo próprio autor deste trabalho) de alguns dos resultados obtidos pela solução desenvolvida.

Com esse propósito, ao todo foram avaliadas manualmente vinte resultados, obtidos a partir de sentenças, selecionadas de forma aleatória, dos relatórios de deliberações (repositório N1), que serviram de entrada para o algoritmo de pesquisa de similaridade – adotando-se como referência o repositório N2 (que armazena os votos das deliberações).

Anota-se que o estabelecimento do quantitativo de sentenças avaliadas não seguiu critério estatístico, tendo sido determinado de forma empírica: assumiu-se que o quantitativo de vinte sentenças foi suficiente para detectar um padrão de resposta e eficiência da solução, pois na análise da referida amostragem não foi percebida relevante variação nas características de resposta da solução sob exame.

A tabela 5 traz alguns exemplos coletados das amostras avaliadas, sendo apresentada na coluna mais à esquerda a sentença fornecida como entrada; a qual se segue a coluna que registra um ou dois dos respectivos resultados mais relevantes retornados pela solução e; por fim, na coluna mais à direita, apresenta-se um resumo da avaliação subjetiva dos resultados obtidos:

Tabela 5 – Exemplos e apontamentos relacionados à validação manual dos resultados da solução desenvolvida

Sentença de entrada (oriunda de N1)	Sentenças, pré-processadas, de saída do algoritmo de busca de similaridade (oriunda de N2)	Apontamentos sobre os resultados
"fase projeto básico necessidade bem caracterizar depósitos solos moles identificando extensões espessuras propriedades geotécnicas projeto básico dados geotécnicos quantidade qualidade permitem soluções projeto bem estudadas econômicas razões todos esforços investigação deverão concentrados nesta fase"	1ª sentença mais similar: "ausência equipe técnica equipamentos realização sondagens levantamentos topográficos arguida razões justificativa presta demonstrar regularidade projeto básico verdade causa potencial deficiência tipologia obra canalização cursos requer estudos preliminares sondagens levantamentos topográficos leito perfeita caracterização (...) motivo legislação regente veda tal prática" (1ª sentença mais similar) 2ª sentença mais similar: "recomendar ministério desenvolvimento regional fundamento art 11 315 2020 art 250 inciso iii elabore normativo próprio especificações técnicas mínimas serem observadas estudos geológicos geotécnicos projetos obras hídricas devendo observar melhores práticas constantes literatura técnica nacional internacional sobre assunto forma assegurar parâmetros qualidade adequados estudos grifos acrescidos"	Vários elementos das sentenças atestam a similaridade tais como (i) trata-se da fase de projeto básico; (ii) a irregularidade central diz respeito a estimativa referente solos; e (iii) causa da situação deve-se a falhas de levantamentos (sejam "estudos", "melhores práticas" ou "sondagens")
"adoção índice usual exigência capacidade financeira superior permitida lei licitações contratos fins verificação qualificação interessados participar certame licitatório afronta art 31 3º 5º lei contraria disposto acórdão câmara min aroldo cedraz min aroldo cedraz min josé múcio"	1ª sentença mais similar: "adequada proposta dar ciência prefeitura municipal adoção índice usual exigência capacidade financeira superior permitida lei licitações contratos fins verificação qualificação interessados participar certame licitatório forma índice denominado capacidade financeira disponível cfd afronta art 31 3º 5º lei contraria disposto acórdãos câmara min aroldo cedraz min aroldo cedraz min josé múcio sendo adequada adoção outra metodologia avaliação diminuição capacidade operativa absorção disponibilidade financeira previstas art 31 4º lei"	Houve similaridade entre os textos de entrada e saída atestada pelas (i) irregularidade central é a exigência de capacidade financeira; e (ii) são citação de legislações e deliberações. Ressalta-se que a sentença de saída fornece a importante informação de que para a situação sob análise a propositura foi o encaminhamento "dar ciência" (que se adota para situações de menor gravidade)
"dessa forma refere lote 4 percentual suficiente afirmar superestimativa quantitativos vista imprecisões inerentes tipo levantamento dados pequena extensão lote 4 utilizada amostra especialmente valor definitivo podendo ter sido alterado medições posteriores analisadas"	1ª sentença mais similar: 'desse modo após cuidadoso extenso exame novos dados unidade técnica verificou fato quantitativos terraplenagem projeto executivo 2011 superestimados adotando referência quantitativos calculados base informações topográficas constatação confirmada lotes 2 3 4 descartada lote 1'	Houve similaridade entre os textos de entrada e saída atestada pelas (i) irregularidade central é a superestimativa de quantitativos de serviços; e (ii) a origem do problema diz respeito a serviços em área aberta (inferido por "pequena extensão lote 4" no texto de entrada e "informações topográficas" no texto de saída.
"desclassificação licitante mpe engenharia serviços"	"dois achados auditoria reportados i restrição competitividade decorrente exigência editalícia licitante comprovasse propriedade usina asfalto contrato compromissório empresa fornecedora asfalto condição necessária habilitação licitantes anexo i edital itens peça 21 pp 32 36 ii excesso formalismo desclassificação licitante tecnicamente habilitada certame"	Interessa observar que a solução partiu, neste caso de uma sentença bem sucinta e retornou em um caso em concreto com descrição mais extensa – o que pode ser bastante útil para rotina de análises técnicas.

Fonte: elaboração própria

Ressalta-se que na tabela 5 foram apresentados apenas alguns resultados que, de fato, na avaliação subjetiva realizada, atenderiam aos objetivos para os quais a solução foi desenhada. Optou-se por trazer à referida tabela os exemplos de textos que versam sobre conteúdos de mais

fácil apreensão para não especialistas nos assuntos e jargões associados à atuação do TCU.

Ainda que a análise apresentada careça de rigor científico (pois não atende propriamente aos requisitos de reprodutibilidade e objetividade), conforme considerações de cunho subjetivo anotadas na última coluna da tabela 5, pode-se assumir que, de forma geral, a solução logrou sucesso em retornar, na maioria dos casos, elementos textuais com potencial de contribuir para análise de questões técnicas rotineiramente avaliadas pelas unidades técnicas do TCU.

5 CONCLUSÕES

A presente pesquisa voltou-se ao processo de recuperação de informação. Diante do enorme volume de documentos e informações (não estruturadas) objeto de estudo, considerou-se as técnicas de Processamento de Linguagem Natural (PLN) como o alicerce para facilitar o acesso, de forma eficiente, a trechos de documentos técnicos potencialmente úteis para o exame de questões abrangidas nos processos de controle externo do TCU.

Em essência, a solução desenvolvida neste trabalho tem por objetivo o auxílio à escrita de análises de processos do TCU, por meio do resgate de argumentos, critérios e encaminhamentos de deliberações anteriores relacionadas a determinado assunto submetido ao exame do referido Tribunal.

O capítulo que trata da metodologia detalha a solução desenvolvida, que se baseia na similaridade de cossenos de textos vetorizados para obtenção de trechos de deliberações do TCU correlacionados ao assunto de uma dada sentença fornecida como entrada. Já conforme destacado no capítulo dedicado à avaliação da solução desenvolvida, não foi possível identificar na literatura acadêmica uma metodologia sedimentada que propiciasse a validação dos resultados obtidos com estrito rigor científico.

Desta forma, foram utilizados procedimentos de avaliação voltados às especificidades da solução implementada, com base nos quais foi possível obter indicativos de que os conteúdos retornados pelo algoritmo de recuperação de informação condisseram, em grau satisfatório, aos assuntos principais das sentenças fornecidas como entrada. Esses indicativos basearam-se (i) no cômputo de ocorrências em que houve correspondência entre a deliberação das sentenças selecionadas por sorteio e dos respectivos resultados retornados; e (ii) avaliação subjetiva do conteúdo dos resultados de simulação realizadas.

Com base nos esforços teóricos e práticos realizados, foi possível inferir que a ideia geral da solução proposta neste trabalho, caso avançasse aos estágios de produção e de aprimoramento baseado em *feedbacks* dos usuários, tem grande potencial de contribuir para a ampliação do conhecimento sobre a produção de textos e a utilização de critérios técnicos voltados à análise de questões recorrentes no âmbito do TCU – o que importaria impactos positivos na transparência e isonomia da atuação desse órgão de controle.

Assim, como **trabalhos futuros**, vislumbra-se uma ferramenta que possibilitasse a utilização de repositório adicional constituído a partir de peças técnicas elaboradas especificamente pelo usuário da solução – personalizando a funcionalidade para permitir o resgate de textos elaborados pelo próprio usuário (com grande possibilidade de reaproveitamento em processos distintos com mínimas adaptações).

Outro desdobramento seria uma funcionalidade que possibilitasse opção de pesquisa de textos nos repositórios de deliberações com o recurso de exclusão de entidades nomeadas (com

auxílio de algoritmo de *Named-Entity Recognition*)[1].A intenção, neste caso, seria obter maior precisão dos resultados da solução de acordo com as necessidades específicas do usuário. Na prática, constata-se, por exemplo, que a informação desejada por um usuário pode estar restrita a deliberações dirigidas a determinado órgão ou legislação (tipos de entidades), mas noutros caso a consulta requerida pode dizer respeito apenas a determinada conduta ou situação (que, a princípio, independe do órgão jurisdicionado ou da legislação envolvida).

Vislumbra-se ainda integração desta solução com um algoritmo baseado em *Retrieval-Augmented Generation* (RAG)[2], com o intuito de gerar respostas adaptadas às especificidades do texto da análise em elaboração pelo usuário. A princípio, entretanto, reconhece-se a técnica RAG apenas de forma complementar (não substitutiva) da solução desenvolvida neste trabalho; pois entende-se que, não raras vezes, é imperioso conhecer os exatos termos da deliberação original ou, ainda, faz-se pertinente o uso de transcrições das deliberações para fundamentação do texto em elaboração.

[1] O recurso *Named-Entity Recognition* (NER), ou reconhecimento de entidades mencionadas (REM) em protuguês, é uma tarefa classificada como estração da infomração pode identificar e categorizar entidades em textos não estruturado, como, por exemplo, pessoas, lugares, organizações e quantidades.

[2] *Retrieval-Augmented Generation* (RAG) é o processo de otimizar a saída de um grande modelo de linguagem, de forma que ele faça referência a uma base de conhecimento confiável fora das suas fontes de dados de treinamento antes de gerar uma resposta. Grandes modelos de linguagem (LLMs) são treinados em grandes volumes de dados e usam bilhões de parâmetros para gerar resultados originais para tarefas como responder a perguntas, traduzir idiomas e concluir frases. A RAG estende os já poderosos recursos dos LLMs para domínios específicos ou para a base de conhecimento interna de uma organização sem a necessidade de treinar novamente o modelo. É uma abordagem econômica para melhorar a produção do LLM, de forma que ele permaneça relevante, preciso e útil em vários contextos. Fonte: O que é RAG (Retrieval-Augmented Generation), disponível em https://aws.amazon.com/pt/what-is/retrieval-augmented-generation/

REFERÊNCIAS

ACKOFF, R. L. From data to wisdom. **Journal of applied systems analysis**, 1989.

ARAMPATZIS, A. *et al.* An evaluation of linguistically-motivated indexing schemes. *In*: SIDNEY SUSSEX COLLEGE CAMBRIDGE. **Proceedings of the 22nd BCS-IRSG Colloquium on IR Research**. [*S.l.: s.n.*], 2000. p. 34–45.

ARAÚJO, R. D. *et al.* Tópicos especiais em sistemas de informaçao: Minicursos sbsi 2022. **Sociedade Brasileira de Computação**, 2022.

AWS. **O que é inteligência artificial geral?** Acessado em 12 de Abril de 2024. Disponível em: https://aws.amazon.com/pt/what-is/artificial-general-intelligence/.

AWS. **O que é RAG (Retrieval-Augmented Generation)**. Acessado em 2 de Setembro de 2024. Disponível em: https://aws.amazon.com/pt/what-is/retrieval-augmented-generation/.

BAEZA-YATES, R. Modern information retrieval. **Addison Wesley google schola**, v. 2, p. 127–136, 1999.

BARNES, J. **Microsoft Azure essentials Azure machine learning**. [*S.l.: s.n.*]: Microsoft Press, 2015.

BRASIL. Constituição da república federativa do brasil promulgada em 5 de outubro de 1988: atualizada até a emenda constitucional n. 48, de 10-8-2005 :. Saraiva„ 1988. Disponível em: http://www.planalto.gov.br/ccivil_03/constituicao/constituicao.htm.

BRASIL. Lei nº. 8.666, de 21 de junho de 1993. **Diário Oficial da República Federativa do Brasil**, Brasília, DF, 1993. Disponível em: http://www.planalto.gov.br/ccivil_03/LEIS/L8666cons.htm.

BRASIL. Lei complementar 101, de 4 de maio de 2000. **Diário Oficial da República Federativa do Brasil**, Brasília, DF, 2000. Disponível em: http://www.planalto.gov.br/ccivil_03/_Ato2011-2014/2012/Lei/L12651.htm.

BRASIL. Lei nº. 13.303, de 30 de junho de 2016. **Diário Oficial da República Federativa do Brasil**, Brasília, DF, 2016. Disponível em: https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2016/lei/l13303.htm.

BRASIL. Lei nº. 14.133, de 1º de abril de 2021. **Diário Oficial da República Federativa do Brasil**, Brasília, DF, 2021. Disponível em: https://www.planalto.gov.br/ccivil_03/_ato2019-2022/2021/lei/l14133.htm.

BRIET, S. Qu'est-ce que la documentation. **Paris, Edit**, 1951.

BUCKLAND, M. **Information and information systems**. [*S.l.: s.n.*]: Bloomsbury Publishing USA, 1991.

BURNHAM, T. F. *et al.* Aprendizagem organizacional e gestão do conhecimento. 2005.

CASELI, H. d. M.; NUNES, M. d. G. V. Processamento de linguagem natural: conceitos, técnicas e aplicações em português. 2023.

CASELI HELENA DE MEDEIROS E NUNES, M. d. G. V.; PAGANO, A. S. O que é pln? **Processamento de linguagem natural: conceitos, técnicas e aplicações em português**, 2023.

CORTES, M. R. P. A. Arquivo público e informação: acesso à informação nos arquivos públicos estaduais do brasil. Universidade Federal de Minas Gerais, 1996.

DEVLIN, J. *et al.* Bert: Pre-training of deep bidirectional transformers for language understanding. **arXiv preprint arXiv:1810.04805**, 2018.

DEVLIN, J. *et al.* **BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding**. 2019.

DIJK van. **Van Dijk. La ciencia del texto: un enfoque interdisciplinario**. [*S.l.: s.n.*]: Barcelona, Buenos Aires: Paidós, 1987.

EBECKEN, N. F. F. *et al.* Mineração de textos. **Sistemas inteligentes: fundamentos e aplicações. São Carlos: Manole**, p. 337–370, 2003.

EDDISON, L. **Deep Learning: A Technical Approach To Artificial Intelligence For Beginners**. [*S.l.: s.n.*]: CreateSpace Independent Publishing Platform, 2017.

FERNANDES, A. M. d. R. Inteligência artificial: noções gerais. **Florianópolis, SC: VisualBooks Editora**, 2003.

FERNEDA, E. **Introdução aos modelos computacionais de recuperação de informação**. [*S.l.: s.n.*]: Ciência moderna, 2012.

FIRTH, J. R. The technique of semantics. **Transactions of the philological society**, Wiley Online Library, v. 34, n. 1, p. 36–73, 1935.

FURTADO, J. S. Informação e organização. **Ciência da Informação**, v. 11, n. 1, 1982.

GOLDBERG, Y. **Neural network methods for natural language processing**. [*S.l.: s.n.*]: Springer Nature, 2022.

GOMEZ-PEREZ, J. M. *et al.* Understanding word embeddings and language models. **A Practical Guide to Hybrid Natural Language Processing: Combining Neural Models and Knowledge Graphs for NLP**, Springer, p. 17–31, 2020.

HAN, B.; SRIHARI, R. Text retrieval using object vectors. **Department of Computer Science and Engineering, State University of New York at Buffalo**, 1999.

HARRIS, Z. S. Distributional structure. **Word**, Taylor & Francis, v. 10, n. 2-3, p. 146–162, 1954.

HAYES, S. C. Behavioral philosophy in the late 1980's. **Theoretical & Philosophical Psychology**, Division 24 of the American Psychological Association, the Division of ..., v. 6, n. 1, p. 39, 1986.

HEARST, M. A. Untangling text data mining. *In*: **Proceedings of the 37th Annual meeting of the Association for Computational Linguistics**. [*S.l.: s.n.*], 1999. p. 3–10.

IBM. **O que é aprendizado supervisionado?** Acessado em 12 de Abril de 2024. Disponível em: https://www.ibm.com/br-pt/topics/supervised-learning.

IBM. **What is text mining?** Acessado em 11 de Abril de 2024. Disponível em: https://www.ibm.com/topics/text-mining.

JURAFSKY, D.; MARTIN, J. H. **Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition**. 2008.

KAO, A.; POTEET, S. R. **Natural language processing and text mining**. [*S.l.: s.n.*]: Springer Science & Business Media, 2007.

LIU, Q.; KUSNER, M. J.; BLUNSOM, P. A survey on contextual embeddings. **arXiv preprint arXiv:2003.07278**, 2020.

LIU, Y. *et al.* **RoBERTa: A Robustly Optimized BERT Pretraining Approach**. 2019.

LOPER, E. E. D. **Applying semantic relation extraction to information retrieval**. 2000. Tese (Doutorado) — Massachusetts Institute of Technology, 2000.

LÓPEZ, M. L. *et al.* Problemas y métodos en el análisis de preposiciones. Biblioteca Hernán Malo González, 1972.

LOVINS, J. B. Development of a stemming algorithm. **Mech. Transl. Comput. Linguistics**, v. 11, n. 1-2, p. 22–31, 1968.

MACHADO, V. P. **Inteligência Artificial. 2010**. 2020. Disponível em Disponível em: https://docplayer.com.br/190820270-Inteligencia-artificial-vinicius-ponte-machado.html. Acesso em: 20 out. 2020.

MAIMON, O.; ROKACH, L. Introduction to knowledge discovery and data mining. *In*: **Data mining and knowledge discovery handbook**. [*S.l.: s.n.*]: Springer, 2010.

MCCARTHY, J. *et al.* A proposal for the dartmouth summer research project on artificial intelligence, august 31, 1955. **AI magazine**, v. 27, n. 4, 2006.

MEIRELLES, H. L. *et al.* **Direito administrativo brasileiro**. [*S.l.: s.n.*]: Revista dos Tribunais, 1991.

MIKOLOV, T. *et al.* Efficient estimation of word representations in vector space. **arXiv preprint arXiv:1301.3781**, 2013.

NAUFEL, J. Novo dicionário jurídico brasileiro. rev. **Atual. Ampli**, v. 3, 1965.

OECD, P. Oecd glossary of statistical terms. Paris (France) OECD, 2008.

OLIVEIRA, H. C. de; CARVALHO, C. L. de. Gestão e representação do conhecimento. 2008.

PASSOS, E. J. L. O controle da informação jurídica no brasil: a contribuição do senado federal. **Ciência da informação**, v. 23, n. 3, 1994.

PASSOS, E. L.; BARROS, L. V. **Fontes de informação para pesquisa em direito**. [*S.l.: s.n.*]: Briquet de Lemos Livros, 2009.

PENNINGTON, J.; SOCHER, R.; MANNING, C. D. Glove: Global vectors for word representation. *In*: **Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)**. [*S.l.: s.n.*], 2014.

PETERS, M. E. *et al.* Deep contextualized word representations. *In*: WALKER, M.; JI, H.; STENT, A. (ed.). **Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)**. New Orleans, Louisiana: Association for Computational Linguistics, 2018. Disponível em: https://aclanthology.org/N18-1202.

PIETSCH, E. **Técnicas modernas de Documentación**. [*S.l.: s.n.*]: Centro de información y documentación, Patronato de investigación científica ..., 1966.

PORTER, M. F. An algorithm for suffix stripping. **Program**, MCB UP Ltd, v. 14, n. 3, p. 130–137, 1980.

RADFORD, A. *et al.* Improving language understanding by generative pre-training. OpenAI, 2018.

REALE, M. **Lições preliminares de direito**. [*S.l.: s.n.*]: Saraiva, 2001.

REIMERS, N.; GUREVYCH, I. **Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks**. 2019.

RIBEIRO, J. V. A. **Exploração de informações contextuais para enriquecimento semântico em representações de textos**. 2018. Tese (Doutorado) — Universidade de São Paulo, 2018.

ROTHMAN, D. **Transformers for Natural Language Processing: Build innovative deep neural network architectures for NLP with Python, PyTorch, TensorFlow, BERT, RoBERTa, and more**. [*S.l.: s.n.*]: Packt Publishing Ltd, 2021.

SALTON, G. The smart system. **Retrieval Results and Future Plans**, Prentice-Hall, v. 260, 1971.

SANH, V. *et al.* **DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter**. 2020.

SANTOS, L. P. dos; TORRES, R. P. A gestão documental enquanto ativo impulsionador da eficiência, transparência e responsabilidade do poder judiciário. **Sistema e-Revista CNJ**, v. 3, n. 2, p. 56–66, 2019.

SIEG, A. From pre-trained word embeddings to pre-trained language models—focus on bert. **Towards Data Science, Online: https://towardsdatascience. com/from-pre-trained-word-embeddings-to-pre-trained-language-models-focus-o n-bert-343815627598**, 2019.

SILVA, D. P. Vocabulário jurídico. **Rio de janeiro: Forense**, v. 2, p. 417, 2004.

SILVA, M. de A. O pré-processamento em mineração de dados como método de suporte à modelagem algorítmica.

SOUZA, S. T. de. A caracterização do documento jurídico para a organização da informação. Universidade Federal de Minas Gerais, 2013.

VASWANI, A. *et al.* Attention is all you need. **Advances in neural information processing systems**, v. 30, 2017.

VIEIRA, S. B. **La recuperación automática de información jurídica: metodología de análisis lógico-sintáctico para la lengua portuguesa**. 1995. Tese (Doutorado) — Universidad Complutense de Madrid, 1995.

VOORHEES, E. M. Using wordnet to disambiguate word senses for text retrieval. *In*: **Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval**. [*S.l.: s.n.*], 1993. p. 171–180.

YANG, J. *et al.* Harnessing the power of llms in practice: A survey on chatgpt and beyond. 2023.