Universidade Estadual de Londrina

KELI RODRIGUES DO AMARAL BENIN

PROCESSAMENTO DE LINGUAGEM NATURAL E A CIÊNCIA DA INFORMAÇÃO: INTER-RELAÇÕES E CONTRIBUIÇÕES

LONDRINA

2023

KELI RODRIGUES DO AMARAL BENIN

PROCESSAMENTO DE LINGUAGEM NATURAL E A CIÊNCIA DA INFORMAÇÃO: INTER-RELAÇÕES E CONTRIBUIÇÕES

Dissertação apresentada ao Programa de Pós-graduação em Ciência da Informação da Universidade Estadual de Londrina, como requisito parcial à obtenção do título de Mestre em Ciência da Informação.

Linha de pesquisa: Organização e Representação da Informação e do Conhecimento.

Orientador: Prof. Dr. Rogério Aparecido Sá Ramalho.

LONDRINA

2023

Ficha de identificação da obra elaborada pelo autor, através do Programa de Geração Automática do Sistema de Bibliotecas da UEL

B467p Benin, Keli Rodrigues do Amaral.
Processamento de Linguagem Natural e a Ciência da Informação : inter-relações e contribuições / Keli Rodrigues do Amaral Benin. - Londrina, 2023.
103 f. : il.

Orientador: Rogério Aparecido Sá Ramalho.
Dissertação (Mestrado em Ciência da Informação) - Universidade Estadual de Londrina, Centro de Educação Comunicação e Artes, Programa de Pós-Graduação em Ciência da Informação, 2023.
Inclui bibliografia.

1. Processamento da Linguagem Natural (PLN) - Tese. 2. Ciência da Informação - Tese. 3. Linguística - Tese. 4. Terminologia - Tese. I. Ramalho, Rogério Aparecido Sá. II. Universidade Estadual de Londrina. Centro de Educação Comunicação e Artes. Programa de Pós-Graduação em Ciência da Informação. III. Título.

CDU 02

KELI RODRIGUES DO AMARAL BENIN

PROCESSAMENTO DE LINGUAGEM NATURAL E A CIÊNCIA DA INFORMAÇÃO: INTER-RELAÇÕES E CONTRIBUIÇÕES

Dissertação apresentada ao Programa de Pós-graduação em Ciência da Informação da Universidade Estadual de Londrina, como requisito parcial à obtenção do título de Mestre em Ciência da Informação. Linha de pesquisa: Organização e Representação da Informação e do Conhecimento.

BANCA EXAMINADORA

Prof. Dr. Rogério Aparecido Sá Ramalho.
Universidade Estadual de Londrina – UEL

Prof. Dra. Brígida Maria Nogueira Cervantes
Universidade Estadual de Londrina – UEL

Prof. Dra. Barbara Coelho Neves
Universidade Federal da Bahia – UFBA

Londrina, 05 de Junho de 2023.

AGRADECIMENTOS

Agradeço primeiramente a Deus, que sempre me ama e me fortalece. Obrigada meu Deus, pois o Senhor é bom o tempo todo, obrigada por me cuidar, me guiar, por me inspirar, por contribuir para que eu sempre faça aquilo que eu preciso fazer, obrigada meu Deus por sempre me orientar e me ilumina com grandes ideias, eu sou muito abençoada por ser sua filha.

Dos meus, quero iniciar falando dos que caminharam antes de mim, acreditaram nos seus sonhos e fizeram suas conquistas. A todos meus antepassados, que me deram força e coragem para eu chegar até aqui. Aos meus pais, que me deram a vida. Pai (José) que meu deu a sua força e me levou para vida; Mãe (Rose) que me nutriu e me protegeu. Vocês são os pais certos para mim e eu sou a filha certa para vocês. Amo e honro vocês.

Aos meus irmãos, amo vocês... e as famílias que constituíram.

Ao meu marido, Luciano pela paciência e força que me deste nessa jornada.

À minha princesa pimentinha Emily, obrigada pela paciência com a mamãe, por todas as horas que não estava disponível para brincar com você, por ser um dos motivos de lutar tanto para evoluir e ser uma pessoa melhor. E você me faz querer ser uma pessoa melhor a cada dia!

Agradeço a Universidade Estadual de Londrina, onde iniciei a minha formação profissional (acadêmica) e hoje retorno a casa para dar continuidade aos estudos. Obrigada pelos excelentes profissionais que possuem e pela oportunidade de avançar na minha carreira acadêmica. Agradeço a todos os professores do PPGCI-UEL, sobretudo aos da linha Organização e Representação da Informação e do Conhecimento.

Ao meu Orientador, Rogério Aparecido Sá Ramalho, pela paciência em me orientar e direcionar o caminho para a conclusão.

Agradeço ainda às professoras que participaram da banca avaliadora desta dissertação, pelo aceite, participação efetiva, cuidado, atenção e carinho com que conduziram o processo, demonstrando que o rigor e a ética são preponderantes no processo de pesquisa, sem deixar de lado o desenvolvimento humano do pesquisador. Meus agradecimentos às professoras e pesquisadoras Brígida e Bárbara.

Agradeço aos meus colegas de trabalho da Biblioteca (José, Carol, Felipe) da Universidade Tecnológica Federal do Paraná (UTFPR) – pôr dá continuidade aos trabalhos sem mim. E a minha chefia imediata (Jean) que me apoiou o meu afastamento.

Agradeço a Universidade Tecnológica Federal do Paraná (UTFPR), por autorizar o meu afastamento das minhas atribuições ao meu cargo.

E por último, mas não menos importante a uma pessoa especial, que me desafiou a tentar o processo seletivo para o mestrado e me auxiliou no projeto, que resultou na minha aprovação no processo de seleção das universidades da UFSC e UEL. Muito Obrigada Emanuelle Torino, pelas suas palavras de incentivo e encorajamento, você é um exemplo de determinação.

Obrigada a todas as pessoas que de alguma forma contribuíram para essa conquista.

A todos, minha gratidão!

Tudo posso naquele que me fortalece. Filipenses: 4.13

Todas as coisas cooperam para o bem
Daqueles que te amam, Jesus. Romanos 8:28

BENIN, Keli Rodrigues do Amaral. **Processamento de Linguagem Natural e a Ciência da Informação: inter-relações e contribuições**. 101f. Dissertação (Mestrado em Ciência da Informação) – Universidade Estadual de Londrina, Londrina, 2023.

RESUMO

Introdução: Nas últimas décadas tem-se observado o aumento na quantidade de informação processada, armazenada e disponibilizada em documentos, principalmente disponibilizados no ambiente digital. Devido ao aumento exponencial da quantidade de informações torna-se inviável a manipulação manual das informações, tornando-se necessário a utilização de mecanismos e ferramentas que possibilitem o processamento automático de informações. A principal abordagem de análise de texto e linguagem por meio computacional é chamada de Processamento de Linguagem Natural (PLN). **Objetivo:** A pesquisa apresenta como objetivo analisar as inter-relações existentes entre o campo do Processamento de Linguagem Natural e a Ciência da Informação, descrevendo as principais contribuições identificadas. **Metodologia**: A pesquisa caracteriza-se como de abordagem qualitativa e descritiva, pois busca identificar e descrever as inter-relações existentes entre o campo do Processamento de Linguagem Natural e a Ciência da Informação. Quanto aos procedimentos metodológicos, a pesquisa pode ser classificada como bibliográfica, fundamentando-se na análise de artigos publicados nas Bases de Dados BRAPCI e SCOPUS, no período de 2012 e 2022, que abordam a inter-relação entre Processamento de Linguagem Natural e Ciência da Informação. Foram selecionados 68 artigos para análise. Após a seleção de artigos, realizou-se a análise de conteúdo de acordo com os ensinamentos de Bardin (2011), onde foram criadas cinco categorias (Artigos de Fundamentação e Conceituais, Pré-processamento de texto; Análise de Semântica e Representação de Texto; Extração de informações e Mineração de texto; Modelagem de tópicos e Classificação de texto) que permitiu realizar análise qualitativa. **Resultados:** Após as análises dos artigos foi possível observar as contribuições do campo de Processamento de Linguagem Natural para a área da Ciência da informação, essas contribuições são mais significativas para a área de recuperação da informação. Referente a área da Ciência da Informação, contribui para o campo de Processamento de Linguagem Natural, com o uso de ontologias, taxonomia, tesauros, dicionários, indexação e organização e representação da informação e do conhecimento. Da mesma forma, foi identificado as contribuições da Linguística e Terminologia para o PLN, com o uso das análises linguísticas e dos termos. **Considerações finais**: Diante disso, foi possível analisar a inter-relações do campo de Processamento de Linguagem Natural com a área da Ciência da Informação, na qual as duas temáticas conseguem usufruir das técnicas e especificidades da outra. Perante o exposto, a Ciência da informação pode e deve usufruir das ferramentas computacionais desenvolvidas no âmbito das pesquisas em PLN, empregando-as nos processos de catalogação e seguido da recuperação nos centros informacionais, assim como na realização de modelos de representação de informação.

Descritores: Processamento da Linguagem Natural (PLN). Ciência da Informação. Linguística. Terminologia. Linguagens Controladas.

BENIN, Keli Rodrigues do Amaral. **Natural Language Processing and Information Science: interrelationships and contributions.** 101f. Dissertation (Master in Information Science) – State University of Londrina, Londrina, 2023.

ABSTRACT

Introduction: In recent decades, there has been an increase in the amount of information processed, stored and made available in documents, mainly made available in the digital environment. Due to the exponential increase in the amount of information, manual manipulation of information becomes impracticable, making it necessary to use mechanisms and tools that enable automatic processing of information. The main approach to text and language analysis by computational means is called Natural Language Processing (NLP). **Objective:** The objective of this research is to analyze the existing interrelationships between the field of Natural Language Processing and Information Science, describing the main contributions identified. **Methodology**: The research is characterized as having a qualitative and descriptive approach, as it seeks to identify and describe the existing interrelationships between the field of Natural Language Processing and Information Science. As for the methodological procedures, the research can be classified as bibliographical, based on the analysis of articles published in the BRAPCI and SCOPUS Databases, in the period 2012 and 2022, which address the interrelationship between Natural Language Processing and Life Science. Information. 68 articles were selected for analysis. After selecting articles, content analysis was carried out according to the teachings of Bardin (2011), where five categories were created (Basic and Conceptual Articles, Text Pre-processing; Semantic Analysis and Text Representation; Information Extraction and Text Mining; Topic Modeling and Text Classification) that allowed for qualitative analysis. **Results:** After analyzing the articles, it was possible to observe the contributions of the field of Natural Language Processing to the area of Information Science, these contributions are more significant for the area of information retrieval. Regarding the area of Information Science, it contributes to the field of Natural Language Processing, with the use of ontologies, taxonomy, thesauri, dictionaries, indexing and organization and representation of information and knowledge. Likewise, the contributions of Linguistics and Terminology to the PLN were identified, with the use of linguistic analyzes and terms. **Final considerations:** Given this, it was possible to analyze the interrelationships of the field of Natural Language Processing with the area of Information Science, in which the two themes are able to take advantage of the techniques and specificities of the other. Given the above, Information Science can and should take advantage of the computational tools developed within the scope of NLP research, using them in the cataloging processes and then retrieval in information centers, as well as in the creation of information representation models.

Descriptors: Natural Language Processing (NLP). Information Science. Linguistics. Terminology. Controlled Languages.

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE ABREVIATURAS E SIGLAS

ABNT	Associação Brasileira de Normas Técnicas
ACC	Assistente de Conhecimento Conceitual
BRAPCI	Base de Dados Referenciais de Artigos de Periódicos em Ciência da Informação
CDD	Classificação Decimal de Dewey
CDU	Classificação Decimal Universal
CI	Ciência da Informação
CC	Ciência da Computação
CMC	Comunicação Mediada por Computador
CoTo	Consensus ou *Trade-Off*
EI	Extração da Informação
EliIE	*Elegibility Criteria Information Extração* (EliIE)
GATE	*General Architecture for Text Engineering*
IA	Inteligência Artificial
IBM	*International Business Machines*
ISO	*International Organization for Standardization*
LE	Ligação de Entidades
LD	Linguagem Documentária
LDA	Alocação Latente de *Dirichlet*
LN	Linguagem Natural
NBR	Norma Brasileira
NLP	*Natural Language Processing*
ORC	Organização e Representação do Conhecimento
PLN	Processamento da Linguagem Natural
TI	Tecnologia da Informação
TIC	Tecnologias da Informação e Comunicação
SN	Sintagma Nominal
SO	Sistemas Operacionais
SVM	Máquina de Vetor de Suporte
TCT	Teoria Comunicativa da Terminologia
TGT	Teoria Geral da Terminologia

SUMÁRIO

1 INTRODUÇÃO

Nas últimas décadas tem-se observado o aumento na quantidade de informação processada, armazenada e disponibilizada em documentos, principalmente acessível no ambiente digital, grande parte do aumento exponencial foi impulsionado pela popularização da Internet e desenvolvimento de novas plataformas digitais, que favorecem maior geração de dados e informações. Segundo Jin *et al*., (2015, p. 60), "[...] o uso extensivo da Internet, Internet das Coisas, Computação em Nuvem e outras tecnologias emergentes de TI fez com que várias fontes de dados aumentassem a uma taxa sem precedentes". Segundo Baeza-Yates; Ribeiro-Neto (1999), julga-se que, grande parte das informações encontradas estava no formato textual, tornando fundamental que os mecanismos de análise e processamento sejam focados nesse tipo de informação

Para contribuir nesses mecanismos de análise e processamento, duas áreas distintas surgem inter-relacionando, quando aplicadas em conjuntos, vem aprimorando diversos campos de pesquisa, e efetivamente, melhorando a qualidade das Tecnologias da Informação e Comunicação (TIC): a Ciência da Informação (CI) e a Ciência da Computação (CC).

Denning *et al.,* (1989, p. 12), definem Ciência da Computação como:

> "[…] o estudo sistemático de processos algorítmicos que descrevem e transferem informação: sua teoria, análise, projeto, eficiência, implementação e aplicação. A questão fundamental de toda a computação é: 'O que pode ser (eficientemente) automatizado?' "

De acordo com essa definição, a Ciência da Computação tem como foco principal os processos que podem ser executados através de um conjunto sequencial de instruções: os algoritmos.

Na introdução do livro intitulado "Ciência da Computação: uma visão abrangente", Brookshear (2013, p. 2), define Ciência da Computação como:

> Uma disciplina que busca construir uma base científica para tópicos como projeto e programação de computadores, processamento de informação, soluções algorítmicas de problemas e o próprio processo algorítmico. Ela fornece a estrutura das aplicações computacionais atuais, bem como a base para a futura infraestrutura de computação.

De fato, a CC trata de algoritmos associados à informação, enquanto a CI se dedica à compreensão da natureza da informação e de seu uso pelos humanos. A CI e a CC são áreas complementares que conduzem às aplicações diversas (SARACEVIC, 1996).

Na visão de Borko (1968), Ciência da Informação é a disciplina que investiga as propriedades e o comportamento informacional, as forças que regem os fluxos de informação, e os significados do processamento da informação para a otimização do acesso e uso.

O autor supracitado esclarece que se trata daquele “corpo de conhecimentos relacionados à produção, coleta, organização, armazenamento, recuperação, interpretação, transmissão, transformação e utilização da informação” (BORKO, 1968, p. 1, tradução nossa).

Robredo (2003), afirma que a interdisciplinaridade da CI não pode restringir apenas no escopo e a abrangência da informação ao campo exclusivo da biblioteconomia e da CI, pois variados estudiosos, pesquisadores e especialistas lidam com a informação de um ponto de vista científico e nas mais variadas abordagens e aplicações, mas sem perder de vista o interesse comum de todos os seus domínios, a entidade informação.

Ainda Lima (2003), aponta as possibilidades de confluência entre a CI e a CC que se concentram nos processos de categorização, indexação, recuperação da informação e interação homem-computador.

Neste contexto, uma das subáreas que mais tem se destacado, em especial dentro da Ciência da Computação, é a Inteligência Artificial (IA). Segundo a autora Gabriel (2022), a IA é a subárea que trabalha com o desenvolvimento de máquinas/computadores com a finalidade de reproduzir a inteligência humana.

Os autores Coneglian e Santarém Segundo (2022), expõem que a IA tem gerado grande impacto em suas aplicações, que estão sendo desenvolvidas por pesquisadores e empresas e utilizadas pelas pessoas em suas rotinas. Os autores ainda afirmam que a IA tem transformado diversas áreas e setores da economia, podendo, quando empregada junto a outras técnicas e teorias, colaborar no aprimoramento do processo da Ciência da Informação.

De acordo com Coneglian, Santarém Segundo (2022, p. 626), dentro da IA, um campo denominado como Processamento de Linguagem Natural (PLN) “tem se destacado pela capacidade de permitir o entendimento do modo como as pessoas se

comunicam, seja de forma oral ou escrita, pelos instrumentos computacionais". O campo de PLN, "quando apoiado por técnicas de aprendizagem de máquina, está tornando o relacionamento entre humanos e máquinas muito mais simples e eficiente" (CONEGLIAN, SANTARÉM SEGUNDO, 2022, p. 626).

Os recentes avanços em PLN por meio de redes neurais têm possibilitado várias pesquisas com tarefas próprias de PLN, como classificação de texto, análise semântica, extração de informação e outros. Essas tarefas de PLN são próprias das atividades da CI, pois entre as várias atividades da CI, pode-se citar a organização e recuperação da informação (FALCÃO, LOPES, SOUZA, 2022).

Um campo que atua nessas tarefas de CI, é a Organização e Representação do Conhecimento (ORC), que busca desenvolver estudos teóricos e metodológicos relativos aos processos de organização do conhecimento, aos conceitos, às suas relações semânticas, além da "[...] busca de teorias e ferramentas para aprimorar as formas de armazenamento e recuperação da informação, entendidas estas como processos que requerem a representação da informação" (FRANCELIN; PINHO, 2011, p. 10).

Dentro desse campo da CI podemos encontrar com o conceito de linguagem documentária, principalmente no que se refere à utilização destas como linguagens de representação de conhecimento.

As linguagens documentárias têm sido aplicadas por unidades de informação para descrever o conteúdo dos documentos. As linguagens documentárias, sejam sistemas de classificação, cabeçalhos de assunto, palavras-chave, lista de descritores ou tesauros, têm o mesmo objetivo e apresentam várias características em comum. Guinchat e Menou (1994, p. 133), complementam que as linguagens documentárias são usadas normalmente no momento de entrada de dados nos sistemas de informação, ou seja, "no tratamento intelectual dos documentos" (análise conceitual e tradução). Ainda segundo os autores, os estudos sobre linguagens documentárias favorecem seus aspectos linguísticos, o que as aproxima das linguagens naturais.

No que tange a linguagem natural, Souza (2005), considera que existem diversas tentativas de se abordar esses processos de representação e recuperação de conhecimento em textos, mas a sua real integração demanda análises concomitantes em diferentes áreas do conhecimento e campos de pesquisa, como a Ciência da Informação, a Linguística, a Terminologia, a Ciência da Computação, a Psicologia Cognitiva, a Comunicação, a Sociologia, dentre outras.

Como mencionado, há sinais evidentes de contribuições de áreas que marcaram e têm influenciado fortemente as pesquisas na área de Processamento de Linguagem Natural: a Linguística, a Ciência da Computação e a Ciência da Informação. Esta massa crítica formada representa uma considerável contribuição à investigação científica, não apenas quantitativamente como qualitativamente (LADEIRA, 2010).

1.1 JUSTIFICATIVA

Justifica-se esse trabalho em três vertentes: a acadêmica, a institucional e a social. A acadêmica, fundamenta-se que a Ciência da Informação está evoluindo e está se aproximando dessa vertente tecnológica e que os profissionais da CI carecem de mais conhecimentos da área da Ciência da Computação, em especial no campo de PLN que utilizam técnicas que auxiliam nos processos de comunicação entre os humanos e as máquinas.

Em relação à vertente institucional, no âmbito do Programa de Pós-Graduação em Ciência da Informação, situa-se a linha de pesquisa Organização e Representação da Informação e do Conhecimento, linha em que esta dissertação se insere, assim, ampliar o escopo de atuação vai ao encontro de seu objetivo, qual seja, promover o conhecimento dos profissionais da informação sobre a organização e representação da informação e do conhecimento no âmbito do Processamento de Linguagem Natural.

E a terceira vertente é a social, espera se que a partir da pesquisa realizada possa trazer contribuições de melhorias nos tipos de serviços oferecidos em unidades de informação, melhor compreensão por parte dos profissionais da informação no que tange às perspectivas e atuações envolvendo o Processamento de Linguagem Natural.

1.2 PROBLEMA DE PESQUISA

De acordo com Nascimento, Martins, Albuquerque (2023), a forma de produção, consumo e organização da informação tem sido alterada e, como resultado a sociedade vivencia uma crescente produtividade informacional, principalmente no ambiente digital, que apoiada pelos avanços oriundos das TIC, tem resultado em um

caos informacional. Diante do excesso de informação das diversas formas de comunicação, os usuários da informação (sejam de bibliotecas, de centros de documentação, dentre outras unidades de informação) atualmente estão cada vez mais exigentes.

Com a consolidação das TIC no cotidiano das pessoas, vem desafiando pesquisadores a compreenderem esse processo e, mais especificamente na área da Ciência da Informação, a encontrarem meios para aperfeiçoar a relação com a informação, elemento cada vez mais valorizado na sociedade (CONEGLIAN, 2020).

Uma forma para esse aperfeiçoamento está nas tecnologias baseadas na Inteligência Artificial, que vem tendo aproximações com a área da Ciência da Informação há alguns anos, com pesquisas sendo realizadas desde o início dos anos 1990 (CUNHA, KOBASHI, 1991). No entanto, com a atual evolução das tecnologias, em especial do campo da Inteligência Artificial, há um cenário bastante favorável para o aprofundamento dessas discussões e de aplicações da IA dentro da Ciência da Informação (CONEGLIAN, 2020).

Na área da Ciência da Informação, a Inteligência Artificial é capaz de auxiliar diversos processos, como a representação, organização e recuperação da informação, permitindo a proposição de novos modelos que utilizam essas novas ferramentas e conceitos. Dentro IA está a subárea, o Processamento de Linguagem Natural, que fornece um ferramental inovador e avançado para a Ciência da Informação executar tarefas como criação, indexação, armazenamento, recuperação e disseminação de informações.

Assim, apresenta-se como problema de pesquisa e necessidade dos profissionais da informação estarem mais familiarizados com as inter-relações existentes entre as práticas identificadas na área da Organização e Representação da Informação e do Conhecimento e os novos desenvolvimentos oriundos do campo de Processamento de Linguagem Natural.

1.3 OBJETIVOS

Mediante o exposto, essa pesquisa tem o objetivo de apresentar as inter-relações existentes entre o campo do Processamento de Linguagem Natural e a Ciência da Informação, descrevendo as principais contribuições identificadas.

Para a consecução do objetivo geral, propõem-se os seguintes objetivos específicos:

i. Identificar as características da Linguística e Terminologia, métodos e processos utilizados na área da Ciência da informação que contribui no Processamento de Linguagem Natural;
ii. Analisar os conceitos de Processamento de Linguagem Natural e verificar como as duas áreas se relacionam;
iii. Refletir a respeito das contribuições da área da CI no contexto do PLN;

1.4 PROCEDIMENTOS METODOLÓGICOS

Segundo Fachin (2017, p. 29), “o método é um instrumento do conhecimento que proporciona aos pesquisadores, [...] orientação geral que facilita planejar uma pesquisa, formular hipóteses, coordenar investigações, realizar experiências e interpretar os resultados”.

A classificação do tipo de pesquisa científica é importante para auxiliar o desenvolvimento de novos métodos científicos, de forma que contribua para a compreensão, análise e discussão de resultados posteriores. Segundo Gerhardt e Silveira (2009), existem diversos tipos de pesquisas, classificadas conforme a abordagem, objetivos da pesquisa e procedimentos técnicos.

No que diz respeito à abordagem da pesquisa é do tipo qualitativa, pois observa e examina quais as possíveis contribuições do campo do Processamento de Linguagem Natural para a Ciência da Informação. Conforme Silva e Menezes (2005, p. 20), “a interpretação dos fenômenos e a atribuição de significados são básicas no processo de pesquisa qualitativa”.

Referente ao objetivo adotou a pesquisa descritiva, pois busca descrever a relação do Processamento de Linguagem Natural com a Ciência da Informação.

Quanto ao procedimento técnico, a pesquisa pode ser classificada em bibliográfica. A pesquisa bibliográfica é desenvolvida a partir de material já elaborado (GIL, 2019), e implica no estudo de artigos, teses, livros e outras publicações usualmente disponibilizadas por editoras e indexadas (WAZLAWICK, 2020).

A fim de descrever mais detalhadamente a metodologia escolhida, dividiremos os procedimentos metodológicos em 2 etapas, exemplificadas na Figura 1 abaixo:

Figura 1 - Etapas dos procedimentos metodológicos

Primeira etapa - Coleta de dados	Segunda etapa - Análise
Levantamento bibliográfico com enfoque nas temáticas do Processamento de Linguagem Natural e da Ciência da Informação; Identificação dos elementos que possibilitam a construção do arcabouço teórico sobre a temática estudada, possibilitando uma melhor compreensão; Busca de artigos nas bases BRAPCI e SCOPUS; Seleção dos artigos com as palavras-chave Processamento de Linguagem Natural, Ciência da Informação, Natural Language Processing OR NLP, Information Science.	Através da análise de conteúdo desenvolvida por Bardin (2011); Resumo e leitura dos artigos que serão selecionados; Definição de categorias para análise dos textos que serão lidos; Informações dos artigos categorizados; Análise de dados que contemplarão as contribuições entre o campo de Processamento de Linguagem Natural e a área da Ciência da Informação.

Fonte: Elaborado pela autora (2022)

Essa divisão foi essencial para que os dados encontrados fossem analisados e processados de maneira mais clara e objetiva. A seguir, descreveremos cada etapa da análise realizada.

Primeira Etapa: Coleta de dados

Nesta etapa foi realizada uma pesquisa bibliográfica a respeito do uso da linguagem na Organização e Representação da Informação e do Conhecimento, também abordando a Linguística e a Terminologia para o entendimento da importância do estudo da linguagem controlada, no qual concluiu, qual era a sua importância e a partir disso foi constatado que atualmente estão abordando esse estudo numa perspectiva automatizada.

Também foi abordada a temática do Processamento de Linguagem Natural e suas principais aplicações.

A pesquisa bibliográfica é fundamental em qualquer tipo de pesquisa. É por meio da mesma que o pesquisador tem contato com inúmeras abordagens sobre o tema pesquisado. Portanto, na pesquisa bibliográfica, o foco foi identificar materiais e

autores que favorecessem um melhor entendimento da temática estudada e trouxessem subsídios para o desenvolvimento da pesquisa.

Na etapa da coleta de dados foram selecionados os artigos nas Bases de Dados em Ciência da Informação, BRAPCI e SCOPUS duas bases de grande importância para pesquisas desenvolvidas por profissionais da informação e áreas correlatas.

Para coleta nas bases de dados, utilizamos como palavras-chave os termos “Processamento de Linguagem Natural” AND “Ciência da Informação” na Base da BRAPCI, foi estabelecido um recorte temporal de janeiro de 2012 a julho de 2022, tendo em vista a contemporaneidade do tema.

Na base de dados SCOPUS utilizamos como palavras-chave os termos “Natural Language Processing” OR “NLP” AND “Information Science”, foi utilizado o mesmo recorte temporal da base da BRAPCI. A coleta de dados ocorreu em julho de 2022. Os critérios de inclusão adotados foram: artigos publicados que contemplassem a temática proposta nos seus títulos, resumos e/ou descritores, disponibilizados na íntegra e nos idiomas português, inglês e espanhol. Consideramos como critérios de exclusão: tese, dissertação ou monografia, artigos com resumo indisponível, os já selecionados na busca em outra base de dados e que não respondessem à questão da pesquisa.

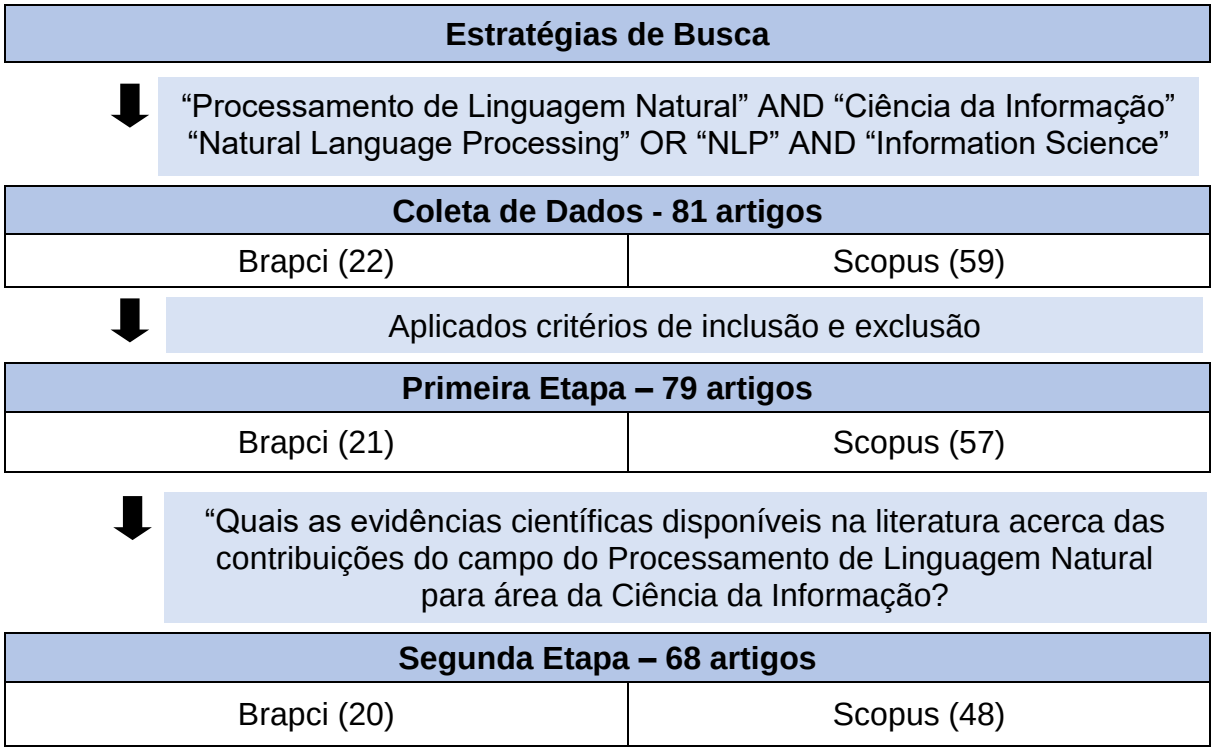
Posteriormente, realizamos a leitura dos títulos dos artigos, resumos, palavras-chave e, quando necessário, a leitura dos textos completos.

A partir da coleta de dados foram encontrados inicialmente um total de 81 artigos científicos. Após a aplicação dos critérios de inclusão e exclusão encontramos um total de 79 artigos. Em seguida, realizamos a leitura flutuante desses artigos, selecionando 68 artigos no total, conforme apresentado no fluxograma da Figura 2:

Com as sintaxes “Processamento de Linguagem Natural”, “Ciência da Informação”, alguns artigos foram desconsiderados: na Base de Dados BRAPCI, dos 22 artigos analisados, 2 foram descartados, pois 1 artigo não possuía conteúdo suficiente sobre PLN e outro artigo foi descartado, pois apresentava apenas o resumo. Já com as sintaxes “Natural Language Processing” OR “NLP”, “Information Science”, na Base de Dados SCOPUS dos 59 artigos recuperados, 11 foram descartados, pois 10 não apresentava relação com PLN e 1 já se encontrava na Base de Dados BRAPCI. Portanto, dos 22 artigos da Base de Dados BRAPCI, foram analisados 20, e

dos 59 artigos da Base de Dados SCOPUS, foram analisados 48, totalizando aqui 68 artigos com as sintaxes mencionadas.

Figura 2 - Fluxograma da busca e seleção dos artigos realizada nas bases de dados.



Fonte: Elaborado pela autora, 2022

Segunda Etapa: Análise de Conteúdo

Com o intuito de analisar artigos com a temática Processamento de Linguagem Natural dentro da Ciência da Informação, nos pautamos no livro Análise de Conteúdo de Bardin (2011), uma obra que pesquisadores utilizam de maneira recorrente devido à sua importância, composta por três fases: organização do material, exploração do material e tratamento dos resultados obtidos e interpretação, descritas a seguir.

Santos F. (2012), fez uma resenha do livro, onde destacou alguns pontos que devem ser considerados pelo pesquisador: na pré-análise ocorre a organização do material que comporá o *corpus* da pesquisa, os documentos são escolhidos, as hipóteses são elaboradas, bem como os indicadores que serão utilizados para a interpretação do texto final. Santos F. (2012), em sua resenha apresenta que esse

contato inicial com a documentação escolhida é chamado por Bardin (2011), de “leitura flutuante”.

Para Bardin (2011), as hipóteses levantadas nessa fase podem ou não ser comprovadas, ou recusadas ao final do estudo. Após a “leitura flutuante”, Bardin (2011), recomenda a utilização de um índice organizado em indicadores e ainda complementa que na fase de exploração do material, os dados são codificados. Nesse processo, os dados são transformados sistematicamente e agregados em unidades.

Como mencionado, na etapa da coleta de dados foram selecionados os artigos nas Bases de Dados BRAPCI e SCOPUS, que fazem parte da primeira fase do método de Bardin.

Na segunda fase, denominada de exploração do material, a “fase de análise propriamente dita, não é mais do que a aplicação sistemática das decisões tomadas” (BARDIN, 2011, p. 131).

Nessa fase, de acordo com Bardin (2011), o objetivo é categorizar os dados selecionados para a análise. Portanto, foi realizada a leitura dos artigos coletados nas bases de dados BRAPCI e SCOPUS. Após essa leitura, foi realizada uma segunda análise, selecionando artigos referentes ao Processamento de Linguagem Natural e Ciência da Informação.

Com o resumo dos artigos selecionados, realizou-se a leitura flutuante, o que possibilitou verificar sob quais aspectos esses temas estão sendo discutidos dentro da Ciência da Informação. Após essa leitura foi possível criar 6 categorias (categorização) de acordo com os conteúdos dos artigos selecionados: Artigos de Fundamentação e Conceituais, Pré-processamento de texto; Análise de Semântica e Representação de Texto; Extração de informações e Mineração de texto; Modelagem de tópicos e Classificação de texto, identificados com informações de autoria, títulos, dentre outros, buscando mais clareza na leitura dos mesmos.

A terceira fase, que recebe o nome de tratamento dos resultados e interpretação, os dados analisados são validados. De acordo com Bardin (2011, p. 133-134, grifos do autor):

> A organização da codificação compreende três escolhas (no caso de uma análise quantitativa e categorial):
> • O recorte: escolha das unidades que pode ser [...] a “palavra” ou a “frase”;
> • A enumeração: escolha das regras de contagem;
> • A classificação e a agregação: escolha das categorias.

A categorização e leitura dos artigos apresentou informações relevantes que serviram de respaldo para analisar e refletir sobre a temática do Processamento de Linguagem Natural no âmbito da Ciência da Informação, o que nos levou ao próximo passo, análise final e apresentação dos resultados.

As informações levantadas na etapa de categorização possibilitaram realizar uma análise final com reflexões acerca das contribuições do campo do Processamento de Linguagem Natural para a Ciência da Informação.

1.5 Estrutura da Dissertação

Com o intuito de facilitar a leitura deste trabalho, estruturamos os capítulos da seguinte maneira:

No capítulo 2, foi realizada uma revisão bibliográfica, com a apresentação de conceitos sobre Linguística e seus tipos de análises linguísticas, os Fundamentos da Terminologia e seus elementos, com o intuito de apresentar a relação com a Ciência da Informação no uso da linguagem controlada.

O capítulo 3, aborda os fundamentos do Processo de Linguagem Natural, aponta as suas principais aplicações, com enfoque em apresentar os conceitos básicos com o intuito de favorecer um melhor entendimento para profissionais da CI.

No capítulo 4 foi efetuada uma análise dos artigos selecionados da área de CI com a temática em Processamento de Linguagem Natural, através de análise de conteúdo proposta por Bardin (2011), que foram selecionados e categorizados, tendo como fonte de pesquisa, as Bases de Dados BRAPCI e SCOPUS. Neste capítulo também foram identificadas e descritas quais as contribuições do campo de Processamento de Linguagem Natural para a área de Ciência da Informação, além de uma reflexão sobre a inter-relação entre as duas áreas.

Por fim, apresentamos as considerações finais a respeito da pesquisa realizada, com discussões a respeito das contribuições de PLN que poderá trazer melhorias para o processamento de informações.

2 O USO DA LINGUAGEM NA CIÊNCIA DA INFORMAÇÃO

Disserta-se neste capítulo a respeito do uso da linguagem na Ciência da Informação. Esse capítulo apresenta os fundamentos da Linguística e seus tipos de análises linguísticas, bem sua contribuição na Ciência da Informação. Também aborda os elementos da Terminologia, que colabora com a linguagem controlada.

2.1 CONTRIBUIÇÕES LINGUÍSTICAS

Segundo Fiori (2003), a Linguística se detém na investigação científica da linguagem verbal humana. Podemos notar que todas as linguagens (verbais ou não-verbais), possuem uma característica importante - são sistemas de signos usados para a comunicação. Saussure a denominou Semiologia; Peirce a chamou de Semiótica. “A Linguística é, portanto, uma parte dessa ciência geral; estuda a principal modalidade dos sistemas sígnicos, as línguas naturais, que são a forma de comunicação mais altamente desenvolvida e de maior uso” (FIORI, 2003, p. 14). Segundo o autor supracitado, a comunicação é essencial para o mundo globalizado de hoje, ou seja, para o mundo de sempre.

A Linguística pode ser considerada uma ciência interdisciplinar, ela conta com a colaboração de vários campos do saber como a Psicanálise, Antropologia, Literatura, Psicolinguística entre outras. Para Lopes (1993, p. 24),

> A linguística é uma ciência interdisciplinar. Ela toma emprestada a sua instrumentação metalinguística dos dados elaborados pela Estatística, pela teoria da Informação, pela Lógica Matemática, etc., e, por outro lado, na sua qualidade de ciência-piloto, ela empresta os métodos e conceitos que elaborou à Psicanálise, à Musicologia, à Antropologia, à Teoria e Crítica Literária, etc.; enfim, ela se dá, como Linguística Aplicada, ao Ensino das Línguas e à Tradução Mecânica.

Isso posto, podemos observar que a Linguística realiza um estudo em conjunto com outras disciplinas, ora como apoio a elas, ora como alicerce delas.

Dentro da Ciência da Informação, segundo Baranow (1983, p 24), além da já “tradicional interface entre Linguística e Ciência da Informação, abrangendo a Morfologia, a Sintaxe e a Semântica, parece que existem novas possibilidades de pesquisas conjuntas, que deveriam ser objeto de um cuidadoso exame crítico”.

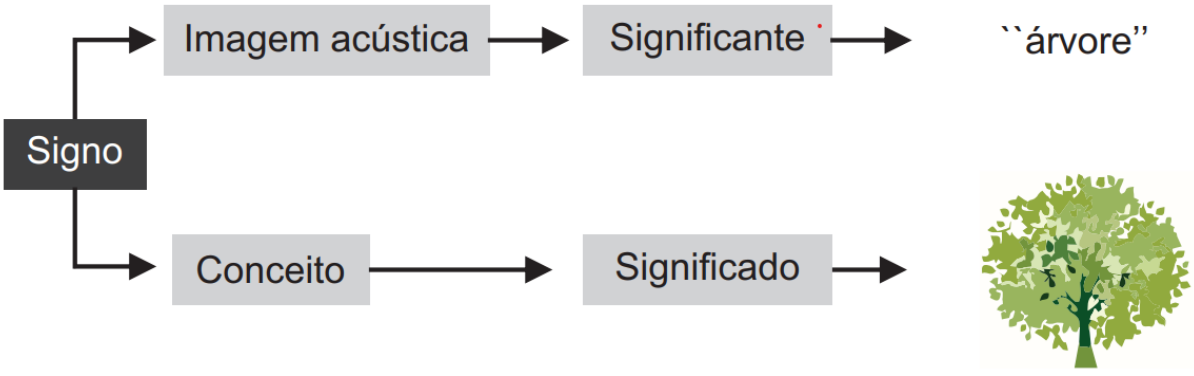
2.1.1 Fundamentos Linguísticos

Em primeiro lugar, devemos estabelecer algumas definições necessárias para compreensão deste subcapítulo, como os conceitos da área de Linguística, iniciamos com a Teoria do Estruturalismo de Ferdinand de Saussure, na qual a Ciência da Informação utiliza a estrutura de organização da linguagem.

Segundo o autor Nascimento (2011, p. 12), “a Linguística é a ciência que estuda toda linguagem verbal ou escrita que faz parte da língua, tendo nela sua matéria de estudo e reflexão”. Para o autor supracitado, a língua é um sistema de elementos que se relacionam entre si e que constituem um código; é uma estrutura organizada, um sistema que se forma de palavras, ou signos linguísticos.

O signo é a união de um conceito com uma imagem acústica (unidos psiquicamente por um vínculo em nossa mente), os quais Saussure chama de significado e significante, respectivamente (FIORIN, 2003), conforme é apresentado na Figura 3.

Figura 3 - Diagrama que ilustra a proposta de Saussure



Fonte: Souza (2018, p.19)

Ao conjunto de significantes ou imagens acústicas atribui-se o nome de plano de expressão, enquanto, o conjunto de significados denomina-se plano do conteúdo (LOPES, 1972). A Semântica adquire espaço no interior da Linguística através da inserção do significado na concepção do signo linguístico.

Para Saussure, o signo linguístico tem duas características principais: a arbitrariedade do signo e a linearidade do significante. Fiorin (2003, p. 76), explica que arbitrariedade do signo é quando a relação entre o significado e o significante é

imotivada, ou seja, “que não há nenhuma relação necessária entre o som e o sentido, que não há nada no significante que lembre o significado”. E já a linearidade do significante “é uma característica das línguas naturais, segundo a qual os signos, uma vez produzidos, dispõem-se uns depois dos outros numa sucessão temporal ou espacial”, por causa dessa característica, não se pode produzir mais de um elemento linguístico de cada vez (FIORIN, 2003, p. 76).

Em relação aos dois eixos de organização da linguagem, o paradigmático e o sintagmático: o eixo paradigmático é o que organiza as relações de oposição (ou/ou, é o eixo da escolha), em que as unidades se substituem (tomo/como); e o eixo sintagmático é aquele que representa as relações de contraste (e+e) em que as unidades se combinam (c+o+m+o= como), conforme o exemplo da Figura 3. Nesse sentido, é que a estrutura da língua estaria sustentada por estas relações de substituição ou combinação de formas (FIORIN, 2003; NASCIMENTO, 2011).

Figura 4 - Eixos da linguagem

Fonte: Lima, Santos, Maimone (2017)

A partir destas concepções linguísticas, podemos observar que a relação da linguística com a Ciência da Informação, pode ser pelo uso desses dois eixos na análise documentária, que se utiliza de métodos e processos para descrever o conteúdo dos documentos.

2.1.2 Tipos de Análises Linguísticas

Interessante mencionar, aqui, que Saussure é conhecido, mundialmente, pelo Curso de Linguística Geral, ele que deu à linguagem uma ciência autônoma, independente. Saussure é referência obrigatória para qualquer teoria linguística. Ele está sempre presente nas mais diversas reflexões a respeito da linguagem (NASCIMENTO, 2011).

A ciência que ele constituiu, a Linguística, tem vertentes que correspondem a diferentes tipos de análise dos diferentes aspectos da língua: sons, palavras, sentenças e discurso nos níveis estruturais, de significado e de uso. Alguns exemplos são apresentados a seguir: análise fonética; análise morfológica; análise semântica; análise pragmática e análise de discurso. Em seguida, são descritas com mais detalhe os tipos de análises:

- **Análise fonética e fonológica**

De acordo com Pinto (2015, p. 6), “a análise fonética consiste no estudo dos sons de uma língua”, isto é, as múltiplas funções dos fonemas, provendo os métodos para a sua descrição, classificação e transcrição. A análise fonológica, no que lhe concerne, estuda o comportamento dos fonemas em uma determinada língua, ou seja, o sistema de som dessa língua (BARROS, ROBIN, 1996). Elas são de maior interesse na implantação de sistema reconhecimento de fala onde é possível o usuário exprimir verbalmente sua busca ou receber alguma forma de resposta audível (FERNEDA, 2003).

No Quadro 1, exemplo de transcrição fonológica automática do sistema de transcrição de grafema para fonema da língua portuguesa.

Quadro 1 - Transcrição de grafema para fonema do português brasileiro

Palavras	Transcrição	Tipo de erro
<pelos>	/'pɛloS/	Transcrição da vogal <e>
<seja>	/'sɛʒa/	
<desde>	/'dɛSde/	
<ter>	/'tɛR/	
<prender>	/preN'dɛR/	
<por>	/pɔR/	Transcrição da vogal <o>
<morreu>	/'mɔʁeu/	
<socorro>	/so'kɔʁo/	
<fossem>	/'fɔseN/	
<força>	/'fɔRsa/	
<México>	/'mɛksiko/	Transcrição da consoante <x>
<aproximando>	/aproksi'maNdo/	
<auxiliar>	/auksili'aR/	
<proximidades>	/proksimi'dadeS/	
<máximo>	/'maksimo/	

Fonte: Santos, Nogueira, Carvalho (2018).

O Quadro 1 apresenta um estudo sobre um sistema automático de transcrição fonológica para o português, utilizando a tecnologia de estados finitos. Nesse quadro também é possível observar os principais erros que se devem à transcrição das vogais <e>, <o> e da consoante <x>.

- **Análise morfológica e lexical**

Na Análise morfológica concentra-se no estudo e classificação de palavras isoladas. Tem o objetivo de dividir o texto em átomos ou *tokens*, fazendo um estudo a cada um desses átomos isoladamente. Os *tokens* podem ser palavras, sinais de pontuação, dígitos entre outros. Aos *tokens* que formam palavras é feita a identificação da sua classe gramatical, o seu lema e o seu radical (PINTO, 2015).

Para Ferneda (2003, p. 83), um exemplo de análise morfológica na recuperação de informação são as técnicas tradicionais de extração de radicais (*stemming*), que visam substituir a variante de uma palavra a uma forma normalizada".

Ainda o autor supracitado, o analisador léxico trata da análise da estrutura e significado da palavra. Um exemplo de análise lexical nos sistemas de recuperação tradicionais é a construção de listas de palavras de pouco valor semântico como artigos e preposições. A análise lexical está relacionada com a criação e uso de vocabulários controlados na indexação de documentos e para a formulação e expansão de expressões de busca (FERNEDA, 2003).

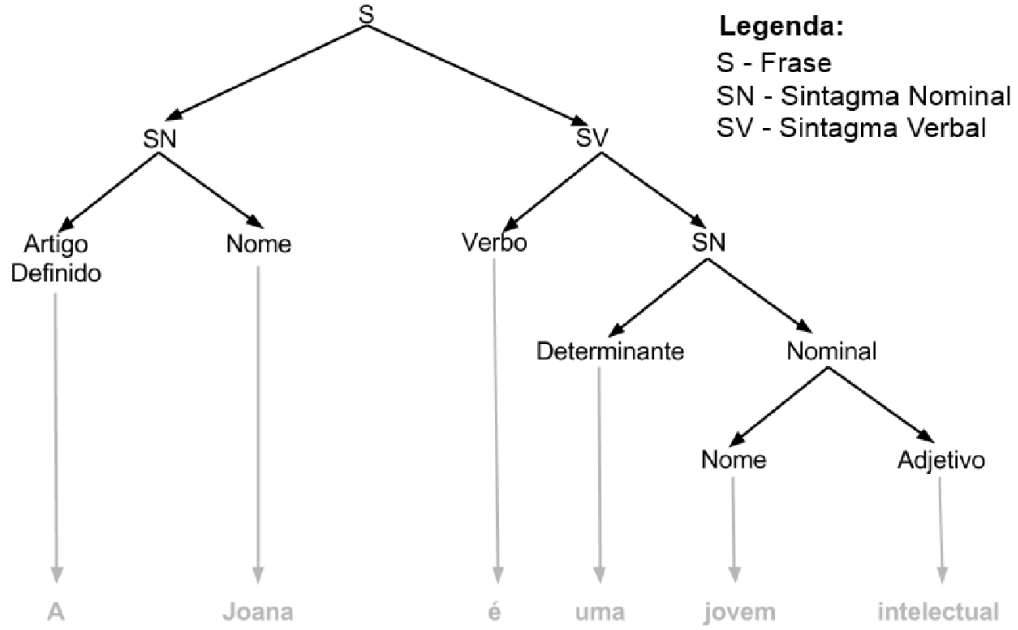
- **Análise sintática**

Para Pinto (2015, p. 6), “a análise sintática, ou *parsing*, propõe-se estudar a relação entre as palavras na frase”. Segundo Dias (2021), as regras gramaticais são aplicadas a categorias e grupos de palavras, não a palavras individuais (como no caso da análise léxica). Para Jurafsky, Martin (2020), a análise sintática basicamente atribui uma estrutura semântica ao texto.

Segundo Crivelli (2011, p. 20), “o analisador sintático cria uma árvore de derivação para cada sentença, mostrando como as palavras estão ligadas entre si”. De acordo o autor supramencionado, é durante a construção da árvore de derivação, feita a averiguação da adequação das sequências de palavras às regras de construção impostas pela linguagem, na composição de frases, períodos ou orações (CRIVELLI, 2011).

Na árvore de derivação, exemplificada na Figura 5 a representação de uma frase passa por explicitar as relações entre os constituintes dela, como, por exemplo, o sintagma adverbial ou sintagma nominal.

Figura 5 - Árvore de derivação para a frase "A Joana é uma jovem intelectual”.



Fonte: Pinto (2015, p. 7)

- **Análise semântica**

A análise semântica é o processo de compreensão do significado e interpretação das palavras, sinais e estrutura da frase, ou seja, análise semântica tem o objetivo de clarificar o significado das palavras num texto (PINTO, 2015; DIAS, 2021). Oliveira (1999), também classifica a semântica em léxica e em gramatical, a semântica léxica busca descrever o sentido através do uso da decomposição semântica das unidades léxicas ou através das redes semânticas que considera como os humanos memorizam e categorizam os conceitos. Já a semântica gramatical tenta buscar o sentido através de uma fórmula lógico-semântica, porém há casos que em uma estrutura pode dar origem a duas representações semânticas, como em “uma professora de capoeira pernambucana” pode referir-se a uma pessoa nascida em Pernambuco ou a uma pessoa que ensina capoeira no estilo Pernambucano.

Segundo Pinto (2015), ao longo dos anos foram criadas diversas representações formais como: lógica de predicados, grafos direcionais e frames semânticos. Na Figura 6, 7 e 8, são exemplificadas as diferentes representações.

Na figura 6, está a representação formal da frase “A biblioteca é um local com livros”, utilizando a lógica de predicados, que denota uma relação entre objetos num determinado contexto.

Figura 6 - Representação formal usando lógica de predicados.

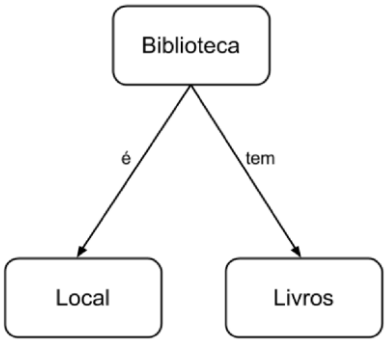
é(biblioteca, local)

tem(biblioteca, livros)

Fonte: Pinto (2015, p. 8)

Na figura 7, está a representação formal da frase “A biblioteca é um local com livros”, usando grafos direcionais, que é uma representação abstrata de um conjunto de objetos e das relações existentes entre eles.

Figura 7 - Representação formal usando grafos direcionais.



Fonte: Pinto (2015, p. 8)

Na figura 8, está a representação do significado da frase “A biblioteca é um local com livros”, usando frames semânticos, que se constitui por um conjunto de frames relacionados por ligações semânticas.

Figura 8 - Representação formal usando frames semânticos.

Biblioteca:
é: local
tem: livros

Fonte: Pinto (2015, p. 8)

- **Análise pragmática**

A análise pragmática é o processo de descobrir o significado de uma frase com base no contexto, circunstâncias externas e expectativas do ouvinte (DIAS, 2021). O contexto inclui compreender como a língua se refere a pessoas e objetos, como o discurso está estruturado e como este é interpretado pelo ouvinte (PINTO, 2015).

Segundo Ferneda (2003), a análise pragmática utiliza conhecimentos externos aos documentos e às buscas do sistema. Este conhecimento pode ser um conhecimento geral do mundo, conhecimento específico para um determinado domínio ou ainda conhecimento sobre as necessidades dos usuários, preferências e objetivos na formulação de uma determinada expressão de busca.

- **Análise de discurso**

Segundo Barros, Robin (1996), a análise de discurso estuda os princípios que governam a produção de sequências estruturadas de frases (discurso – texto ou verbal). Fatores de coesão e coerência do discurso também são abordados aqui. Em outras palavras, a análise de discurso envolve o contexto da comunicação, revelando: quem, para quem e sobre o que se fala.

De acordo com Ferneda (2003, p. 83), para os objetivos da recuperação de informação, a análise de discurso examina a estrutura e os princípios organizacionais de um documento “para entender qual é função específica de uma informação em um documento, por exemplo – é uma conclusão, é uma opinião, uma previsão ou um fato?”.

2.2 ELEMENTOS DA TERMINOLOGIA

A Terminologia é uma disciplina teórica e aplicada que se serve da Linguística, das Ciências da Comunicação, das Ciências Cognitivas, da Ciência da Informação e das especialidades particulares. É um “campo inter e transdisciplinar que envolve a descrição e o ordenamento do conhecimento (nível cognitivo) e a sua transferência (nível comunicacional), e tem como elementos centrais os conceitos e termos” (LARA, 2005, p.1).

> O termo é abordado como uma unidade lexical que adquire um valor especializado, ou terminológico, no léxico dos especialistas de domínio. O conceito, elemento distintivo da Terminologia wüsteriana, passa a ser entendido como um aspecto da unidade terminológica, podendo ser abordado como objeto de estudo da semântica lexical (ALMEIDA, 2021, p. 29).

O foco da Terminologia é a unidade terminológica, ou seja, o termo, considerando os aspectos linguísticos, cognitivos e pragmáticos. Os termos compõem um subconjunto de signos linguísticos correspondente a uma área conceitual e especializada, tendo papel fundamental no desenvolvimento de tesauros, glossários, dicionários, etc. (SOUZA, 2018)

De acordo com Lara (2005, p. 1), termo polissêmico, a Terminologia pode ser definida a partir, pelo menos, de dois aspectos:

> a) Terminologia teórica - conjunto de diretrizes e princípios que governam a compilação, formação de termos e estruturação de campos conceituais (ou nocionais).
> b) Terminologia concreta - conjunto de termos que representam sistemas de conceitos relacionados a uma língua de especialidade ou área de atividade particular.

A Terminologia surge da necessidade de padronizar os termos para facilitar a comunicação entre especialistas.

A Terminologia teve sua origem nos estudos de Wüster, nas primeiras décadas do século XX, que resultaram na Teoria Geral da Terminologia (TGT) publicada como tese e difundida em seus artigos. O objetivo de sua teoria era conseguir uma comunicação inequívoca e sem ambiguidade sobre os temas especializados, uma Terminologia normatizada (ALMEIDA, 2003).

Muitas críticas à TGT surgiram, de acordo com alguns especialistas em Terminologia, a teoria de Wüster não permite descrever adequadamente o léxico especializado (ALMEIDA, 2003).

Segundo Lara (2006) as novas vertentes da Terminologia são socialmente orientadas. Dentre as principais se destacam a Teoria Comunicativa da Terminologia (CABRÉ, 1999), a Socioterminologia (GAUDIN, 1993), a Teoria Sociocognitiva da Terminologia ou Socioontologia, ou, ainda, teoria realista da Terminologia (TEMMERMMAN, 2001) e a Terminologia cultural (DIKI-KIDIRI, 2000). Questionamentos à Terminologia clássica também são feitos por linguistas e psicólogos ligados à Semântica Cognitiva, como Dubois (2001) e Rastier (1995), dentre outros. Nos limites deste trabalho, porém, restringimo-nos a sintetizar a vertente que têm sido mais referidas na literatura terminológica recente.

A Teoria Comunicativa da Terminologia (TCT), proposta por Cabré, propõe ver os termos como unidades linguísticas, enfatizando a função da língua como instrumento de comunicação. A vertente se constitui na confluência entre a teoria do conhecimento, da comunicação e da linguagem" (LARA, 2006, p. 4). A função dos termos para a TCT é dupla: representar e transferir o conhecimento especializado, em graus e modos distintos, como em situações diversas.

2.2.1 Objeto da Terminologia e de Seus Constituintes

Como já foi dito, o termo é o objeto de estudo da Terminologia. Por isso, é necessário que se defina a noção de termo e de seus constituintes.

Segundo Pontes (1997, p. 47), "o termo, que é o objeto de estudo da Terminologia, é basicamente um signo linguístico formado de uma denominação (significante) e um conceito (significado)":

Figura 9 – O termo

$$\mathbf{T}\ (\text{termo}) \;-\; \frac{\text{D (denominação)}}{\text{C (conceito)}} \;-\; \frac{\text{S (significante)}}{\text{S (significado)}}$$

Fonte: Pontes (1997, p. 47)

O que diferencia o termo dos outros signos linguísticos é que sua amplitude semântica se define antes pela relação com o significado do que com o significante. Outra característica do termo é o significado ser definindo na relação com significados pertencentes ao mesmo domínio, que pode ser uma disciplina, uma ciência, uma técnica (PONTES, 1997)

Ainda o autor supracitado, uma terceira característica do termo, é para uma dada noção, haver apenas uma denominação. Tal característica, dentro da Terminologia, se baseia no princípio da univocidade entre denominação e conceito, no qual, se refere à relação entre os termos utilizados para se referir a algo (denominação) e o próprio objeto, ideia ou fenômeno que o termo representa (conceito). A univocidade pressupõe que um termo tenha apenas um significado fixo e inequívoco, ou seja, que haja uma correspondência unívoca entre a palavra e o conceito que ela representa.

De acordo com Pontes (1997, p. 47), uma quarta característica do termo diz respeito ao modo de formação cuja origem se encontra:

> • na especificação de um item lexical da língua comum;
> • na criação neológica, a partir dos múltiplos modos deformação no plano morfológico, morfossintático ou morfossemântico;
> • no recurso a formas perifrásticas ou sintagmáticas mais ou menos complexas.

O termo caracteriza-se, enfim, pelo fato de a homonímia não constituir uma ameaça de ambiguidade, pois o termo se baseia na ligação de um termo a um campo nocional determinado.

Segundo Pontes (1997, p. 47), “a denominação é a forma linguística externa do termo. É, igualmente, o resultado de uma relação estabelecida, seja pelo uso, seja pela criação artificial para representar o conceito”.

A relação que se estabelece, segundo Pontes (1997), entre uma denominação e um conceito é monorreferencial, ou seja, para um dado termo, temos apenas uma denominação. Esta relação é semelhantemente unívoca, quer dizer que para um termo dado, corresponde apenas um conceito.

Para Pontes (1997) verifica-se que o conceito (o mesmo que noção, para ISO), exerce um lugar fundamental na teoria geral da Terminologia, porque é o conteúdo do termo. Dahlberg (*apud* GOMES e CAMPOS, 2019, p. 39), define conceito como uma “unidade do conhecimento, compreendendo afirmações verificáveis sobre

um dado item de referência, representado numa forma verbal". Essa definição pode ser apresentada através do seu triângulo conceitual:

Figura 10 - Triângulo Conceitual de Dahlberg

Fonte: Gomes e Campos (2019, p. 39)

A partir desse modelo formal (modelo classificatório da realidade), é que poderemos estabelecer o que são conceitos. Para Dahlberg, a relação entre características e designação é decisiva para o conhecimento de coisas ou atividades (PONTES, 1997).

O conceito, segundo a Norma ISO 1087-1 (2000), é "[...] unidade de conhecimento constituída por abstração, com base em um conjunto de traços ou características comuns, atribuídas a uma classe de objetos, de relações ou de entidades".

Para a Norma NBR 12676 (ABNT, 1992, p. 1), o conceito é "qualquer unidade de pensamento. O conceito pode ter o seu conteúdo semântico reexpresso pela combinação de outros conceitos, que podem variar de uma língua ou de uma cultura para outra".

2.2.2 Influência da Terminologia na Ciência da Informação

A Terminologia se aplica à comunicação direta, à mediação comunicativa e ao planejamento linguístico. Na área de Documentação, a Terminologia é fundamental para representar o conteúdo dos documentos e para facilitar o acesso a esse conteúdo (DIAS, 2000). Cabré (1995), aponta os tesauros e as classificações como inventários terminológicos organizados de acordo com sua temática e controlados formalmente

As análises terminológicas prestam significativas contribuições às linguagens documentárias na produção de vocabulários controlados, pois:

> (...) a Terminologia objetiva agilizar a comunicação entre especialistas, bem como entre especialistas e público em geral. Assume funções de comunicação e de representação, procura o consenso e propõe formas de controle da diversidade da significação (CERVANTES, FUJITA, RUBI, 2008, p. 214).

Para Dias (2000, p. 91), “a Terminologia representa o conhecimento técnico-científico especializado de forma organizada, por meio de manuais e glossários, e unifica esse conhecimento sob a forma de normas e padrões”. O autor supracitado, ressalta que sem a Terminologia, os especialistas não conseguiriam se comunicar, repassar seus conhecimentos, nem tampouco representar esse conhecimento de forma organizada. Nesse sentido, Cabré (1995), atribui à Terminologia a qualidade de ser a base do pensamento especializado.

A apropriação da Terminologia pela CI na Linguagem Documentária (LD) ocorreu através das normas terminológicas internacionais (ISO 704 e ISO 1087, entre outras) e dos textos teóricos fundamentais., que resultou no aperfeiçoamento das metodologias de construção da LD, visto que as normas terminológicas são, relativamente à norma de elaboração de tesauro (ISO 2788) mais completas, possibilitando uma melhor compreensão dos conceitos, das relações entre os conceitos e da modelagem dos sistemas conceituais. Para o uso da Terminologia concreta - dicionários técnicos especializados, glossários, mapas conceituais, listas de termos com ou sem equivalências, etc. - significou a possibilidade de conferir referência concreta à interpretação dos descritores dos tesauros (LARA, 2006).

Um padrão de aplicação da Terminologia para a documentação é substituir as práticas empíricas de escolha de termos para compor um vocabulário controlado pela consulta a terminologias das áreas do saber ou de atividade, fornecendo metodologias para identificar: os domínios de conhecimento e de atividade; termos e conceitos; as relações entre os conceitos a partir de definições (LARA; LIMA, 2015).

Em vista disso, podemos concluir que a Linguística e a Terminologia têm contribuído para a CI por intermédio da linguagem controlada.

2.3 Linguagens Controladas

Segundo Mendonça (2000), a Ciência da Informação tem uma estreita ligação com a Linguística pela intermediação da Análise Documentária. Ainda a autora destaca que os estudos mais recentes apontam para a tendência da Terminologia e

Análise Documentária, a buscar na próxima década o rigor para as práticas de construção de vocabulários para fins de documentação.

Kobashi (1996), ressalta que a aproximação entre a Análise Documentária e a Linguística Aplicada, justifica-se inicialmente pelas semelhanças existentes entre os processos documentários e a tradução automática, resguardada, obviamente, as especificidades de cada uma delas.

O termo Análise Documentária foi criado por Jean-Claude Gardin para caracterizar as operações semânticas que transformam um texto original em uma ou várias palavras-chave, ou ainda, paráfrases, visando facilitar a representação de 'conteúdos' e a recuperação da informação. Nos trabalhos iniciais de Gardin, a Análise Documentária é definida com um tipo de análise descritiva cuja o objetivo é a de proporcionar uma representação sistemática de certos fatos que se supõem parcialmente ordenados (LARA, 2009).

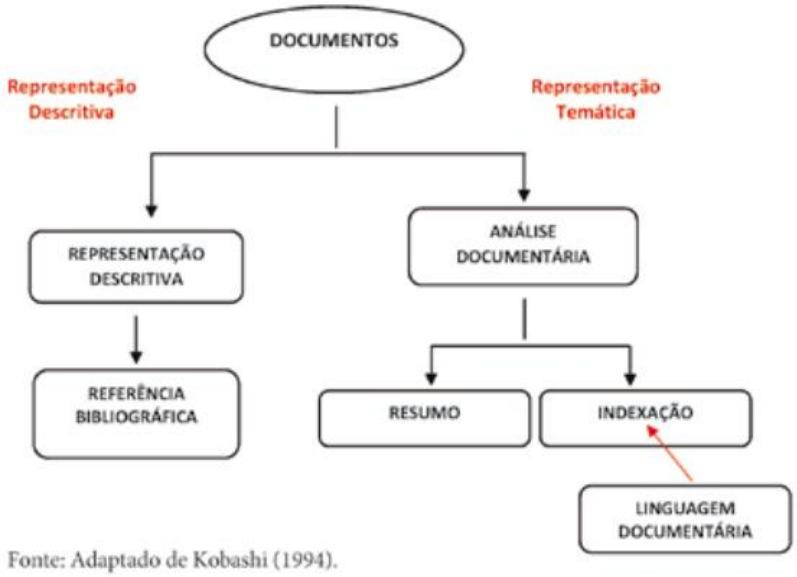
Análise Documentária segundo Gardin (1987, p. 48-49, *apud* KOBASHI, 2008, p. 48), é o "conjunto de procedimentos utilizados para exprimir o conteúdo dos documentos científicos sob formas destinadas a facilitar a sua localização ou consulta"

O objetivo da Análise Documentária é a de excluir o sentido dos textos visando permitir pesquisas retrospectivas da informação a partir de seus conteúdos ou significação (GARDIN, 1973 *apud* LARA, 2009).

A Análise Documentária estabelece relação entre o conceito de representação do resultado das operações de análise e síntese do conteúdo de textos com o objetivo de transferência de informação (LARA, 1999).

Segundo Kobashi (2008), a Análise Documentária pode ser definida como operação com textos: análise e síntese, com o objetivo de obter a distinção entre informação essencial e acessória. Após selecionadas as informações, são estruturadas em novos textos, ditos resumos, ou em símbolos de uma linguagem específica, dita Linguagem Documentária, conforme a Figura 12:

Figura 11 - Tratamento documentário



Fonte: Maimone, Kobashi, Mota (2016, p. 75)

Gardin concebe a operação de Análise Documentária como:

> Uma operação semântica, mesmo que ela não obedeça a uma regra precisa, e que cada organismo, cada analista (…) se limite a ver na ocorrência uma certa regularidade interna fundada mais na experiência ou hábito do que sobre algum procedimento explícito (GARDIN, 1970, *apud* LARA, 2009, p. 28).

Os investimentos metodológicos realizados pela Análise Documentária partiram do princípio que a formalização dos procedimentos nessas operações poderia alterar o quadro empírico de representação no âmbito dos sistemas documentários.

2.3.1 Fundamento de Linguagem Controlada

Inicialmente a documentação trabalhava de forma empírica, coletando termos que iriam constituir uma Linguagem Documentária (LD) a partir da verificação da frequência e ocorrência dos termos na literatura, com os estudos da Linguística, pôde-se verificar que tais linguagens só poderiam efetivamente receber esse nome (Linguagens Documentárias) se funcionassem simultaneamente como instrumentos de significação e de comunicação (LARA, 2005).

Linguagem Documentária, também conhecida como: linguagem controlada ou linguagem de indexação, pode ser definida como um conjunto limitado de termos

autorizados para uso na indexação e busca de documentos, reduzindo substancialmente a diversidade de Terminologia (LOPES, 2000).

Segundo Gardin *et al.,* (1968 apud CINTRA *et al*., 1994, p. 25), são três os elementos básicos de uma LD:

> 1) *Um léxico*, identificado com uma lista de elementos descritores, devidamente filtrados e depurados;
> 2) *Uma rede paradigmática* para traduzir certas relações, essenciais e, geralmente estáveis, entre os descritores. Essa rede, organizada de maneira lógico- semântica, corresponde a uma organização dos descritores numa forma que, *lato sensu*, poderia se chamar de classificatória; e
> 3) *Uma rede sintagmática* destinada a expressar as relações contingentes entre os descritores, relações essas que só são válidas no contexto particular onde aparecem. A construção de sintagmas é feita através de regras sintáticas destinadas a coordenar os termos que dão conta do tema

Entende-se o léxico como um conjunto de palavras de um idioma ou área de especialidade. Enquanto a rede paradigmática é tida como a relação entre as palavras cujo significado é de senso comum entre os especialistas da área. Já a rede sintagmática refere-se às relações que podem ser determinadas entre os termos (CERVANTES, 2009)

As Linguagens Documentárias são linguagens estruturadas e controladas, compostas a partir de princípios e de significados advindos de termos constituintes da linguagem de especialidades e da linguagem natural (LN), com o objetivo de representar para recuperar a informação documentária (BOCCATO, 2009).

Segundo Kobashi (2008), as Linguagens Documentárias são instrumentos privilegiados de mediação que apresentam dupla função: representar o conhecimento; e promover interação entre usuário e conteúdo.

O objetivo da LD é o controle de vocabulário, isto é, controlar a Terminologia de área ou áreas do conhecimento por meio do estabelecimento de um conceito/interpretação, definido aos termos conforme as necessidades de uso do sistema (ALVARES, 2011).

2.3.2 Tipologia das Linguagens Documentárias

Segundo Guimarães (1990 *apud* CERVANTES, 2009), as Linguagens Documentárias podem ser classificadas de acordo com dois critérios: quanto à

ordenação dos conceitos, pré ou pós-coordenadas e quanto a sua forma de apresentação, ordem sistemática ou alfabética. Quanto à ordenação dos conceitos, elas podem ser pré-coordenadas, como os cabeçalhos de assuntos ou pós-coordenadas, como os tesauros. As linguagens pré-coordenadas são aquelas em que o indexador coordena os assuntos quando faz o tratamento da informação. Isto é, o indexador determina quais os assuntos de um documento e procura reuni-los sob formas pelas quais imagina que o usuário irá pesquisar (VALE, 1987).

Assim, ao se indexar um documento que versa sobre os Tratamentos de mastite de bovinos de leite, o indexador, ao pré-coordenar, poderia representar o conteúdo de três maneira:

Bovinos de Leite – Mastite Bovino
Mastite Bovina – Doença – Tratamento
Bovinos – Doenças – Tratamento

Neste caso o usuário só teria acesso a esta obra se pesquisasse exatamente da mesma forma como o seu conteúdo foi representado pelo indexador. O primeiro termo é o que determina a recuperação, o que significa que na pré-coordenação, é necessário o uso de muitas remissivas ou entradas múltiplas para evidenciar todos os conceitos significativos, causando um acréscimo de custo para o sistema na fase da entrada de dados. A pré-coordenação algumas vezes acaba dispersando elementos de conceitos relacionados (VALE, 1987).

Já as linguagens pós-coordenadas são aquelas em que o usuário coordena os assuntos no momento em que busca a informação (VALE, 1987). Assim, ao se indexar a mesma obra, tendo como princípio a pós-coordenação, o indexador representaria separadamente cada assunto, como, por exemplo:

Bovinos de Leite
Mastite Bovina
Bovinos
Doenças
Tratamento

Aqui o indexador permite que o usuário realize a pós-coordenação e chegue até a informação que lhe interessa ao procurar somente um destes descritores, ou ao combinar aleatoriamente os mesmos. A pós-coordenação dispensa a ordem de citação e possibilita múltiplas combinações no momento da pesquisa. Opera preferencialmente com conceitos simples, sendo que estes conceitos podem ter uma ou mais palavras. O uso de conceitos compostos próprios de determinadas áreas auxilia para a especificidade da linguagem, possibilitando melhor exatidão na recuperação (VALE, 1987)

Segundo Lima (1998), quanto a forma de apresentação, as Linguagens Documentárias podem ser classificadas, isto é, apresentam uma ordem sistemática, como a Classificação Decimal de Dewey (CDD), a Classificação Decimal Universal (CDU), a Classificação de Dois Pontos e a Library of Congress, ou alfabéticas como as Listas de Cabeçalhos de Assuntos, o Precis e os Tesauros.

Atualmente possuímos várias ferramentas tecnológicas, que não é possível, mas lidamos com o uso dessa linguagem apenas dessa maneira manual, precisamos automatizar esse processo de utilização de linguagem, para isso o próximo capítulo, trataremos do processamento automático de linguagem natural.

3 PROCESSAMENTO DE LINGUAGEM NATURAL (PLN)

Neste capítulo foram abordados os principais conceitos, aplicações, ferramentas e subcampos do Processamento de Linguagem Natural.

3.1 Fundamentos do Processamento de Linguagem Natural

Segundo as autoras Comarella e Café (2008), dentre os grandes desafios da computação está o de desenvolver-se meios para tornar a comunicação homem-máquina mais natural e intuitiva. Atualmente, busca-se desenvolver programas capazes de “compreender”, mesmo que de forma rudimentar, fragmentos da linguagem humana.

A respeito disso, Silva (2006, p. 104), corrobora:

> É o desafio posto pelo tratamento computacional das línguas naturais e pelo próprio processo de comunicação humano que tem instigado os centros de tecnologia da linguagem humana a investirem significativos recursos teóricos, humanos e materiais na modelagem computacional da linguagem humana, entendida, aqui, como a criação de um modelo computacionalmente tratável do uso do léxico e da gramática de uma língua natural nas diversas situações comunicativas. Nasce, assim, o domínio de estudo conhecido por Processamento Automático de Línguas Naturais (doravante PLN).

O Processamento de Linguagem Natural, ou PLN como é conhecido, é uma área integrante da Inteligência Artificial, em que o aprendizado de máquina e a linguística são amplamente utilizados. Esse campo tem a finalidade de compreender, analisar e gerar a língua natural para os Humanos, para que, eventualmente, seja possível nos dirigirmos a um computador da mesma forma que nos dirigimos a uma pessoa (PINTO, 2015; SANTOS, 2021). Em outras palavras, o Processamento de Linguagem Natural “pode ser definido como a habilidade de um computador em processar a mesma linguagem que os humanos usam no dia a dia” (ROSA, 2011, p.137).

Segundo Chowdhury (2003), o PLN é uma área de pesquisa e de aplicação que explora como os computadores podem ser usados para processar e manipular texto ou discurso em linguagem natural para fazer coisas úteis. As principais áreas que dão subsídios ao processamento da linguagem natural são: a linguística, a semântica, o processamento de sinais e a teoria da comunicação, entre outras. Na

Figura 13, apresenta-se a sistemática elaborada por Silva (2006), para representar os principais recursos teórico-metodológicos que contribuem para o estudo do PLN.

Figura 12 - Recursos teórico-metodológicos para o estudo do PLN

Fonte: adaptado do Silva (2006, p.132).

Segundo Crivelli (2011), os primeiros estudos na área de PLN tiveram início no ano de 1950. Os primeiros sistemas desenvolvidos nos anos 60 já conseguiam responder de forma rudimentar perguntas enviadas pelo usuário sobre muitos assuntos, como por exemplo, matemática e inglês.

O PLN obteve diversos avanços ao longo de meio século de pesquisas. Silva (2006), descreve esses avanços, em termos de sofisticação linguística, na forma apresentada no Quadro 2.

Quadro 2 - Evolução do estudo do PLN

Década	Foco da Investigação	Conquistas
50	Explorações: tradução automática.	▪ sistematização computacional das classes de palavras descritas nos manuais de gramática tradicional; ▪ identificação computacional de constituintes oracionais.
60	Formalizações: novas aplicações e criação de formalismo.	▪ primeiros tratamentos computacionais das gramáticas livres de contexto; ▪ criação dos primeiros analisadores sintáticos; ▪ primeiras formalizações do significado em termos de redes semânticas.
70	Criação do nicho de pesquisa: consolidação do PLN.	▪ implementação de parcelas das primeiras gramáticas e analisadores sintáticos baseados na gramática gerativo-transformacional; ▪ busca de formalização de fatores pragmáticos e discursivos.
80	Busca da precisão: sofisticação dos sistemas.	▪ desenvolvimento de teorias linguísticas motivadas pelos estudos do PLN como, por exemplo, a gramática sintagmática generalizada e a gramática léxico-funcional.
90	Busca da precisão e robustez: sistemas baseados em representações do conhecimento no tratamento estatístico de massa de textos.	▪ Desenvolvimento de projetos de sistemas de PLN complexos que buscam a integração dos vários tipos de conhecimentos linguísticos e extralinguísticos e das estratégias de inferência envolvidos nos processos de produção, manipulação e interpretação de objetos linguísticos para os quais os sistemas são projetados. ▪ ressurgimento da linguística de corpus e do tratamento estatístico de entidades e processos linguísticos.

Fonte: Silva (2006, p.120).

O autor Martins *et al.,* (2020), descreve que a evolução do PLN, deve seu primeiro marco em 1950, pelo matemático, cientista da computação e pesquisador Alan Turing, quando da publicação na ocasião do trabalho *Computing machinery and intelligence*. Em 1954 Turing, destacou-se a experiência de Georgetown, na qual se fez a tradução automática de mais de 60 frases em russo para o inglês (MARTINS *et al*., 2020).

Na década de 1960, é possível indicar sistemas bem-sucedidos de PLN denominados SHRDLU, programas desenvolvidos para a interação humana com termos em inglês. ELIZA foi o primeiro software para simulação de diálogos, empregando pouca informação sobre o pensamento e/ou a emoção humana para a criação do diálogo, também tido como um tipo de chatterbot (MARTINS *et al*., 2020).

Já por volta dos anos 1970 e 1980, diversos programadores iniciaram a escrita do que se chamou “ontologias conceituais”, responsáveis pela estruturação das informações reais em dados que eram compreensíveis para o computador. Ao final da década de 1980, houve o que foi considerada a revolução no Processamento de Linguagem Natural, sobretudo pela introdução de algoritmos de aprendizagem de máquina, ou, como mais conhecida, *machine learning* (MARTINS *et al*., 2020).

Na subseção seguinte foram apresentadas as principais aplicações do PLN no nosso cotidiano.

3.2 Principais Aplicações de PLN

A pesquisa que tem sido feita sobre PLN abrange várias áreas, como a da saúde, de entretenimento e, principalmente, de negócios. Para isso, a máquina interpreta elementos importantes das frases na linguagem humana para retornar respostas adequadas. Em outras palavras, o PLN representa uma abordagem automática para lidar com a linguagem humana, podendo auxiliar em muitas tarefas e na resolução de problemas básicos diários. Martins *et al.,* (2020) especificam algumas aplicações que utilizam PLN, como os descritos a seguir:

Um dos principais usos do PLN são os Assistentes Virtuais Inteligentes, também conhecidos como Assistentes Pessoais Virtuais, são *softwares* projetados para interagir com diversos tipos de usuários em linguagem natural (MATOS, OLIVEIRA, 2021). É um software que auxilia uma pessoa em uma ou mais tarefas, sendo visto como uma criação dotada de inteligência. Existem no mercado inúmeros Assistentes Virtuais Inteligentes, atualmente disponíveis na palma da mão dos usuários em seus smartphones (FUSCHILO, ALENCAR, SCHMITZ, 2019). Alguns exemplos conhecidos são: a Siri, da Apple, a Alexa, da Amazon e a Cortana da Microsoft.

A interação com esses assistentes tem a intenção de requerer o mínimo de esforço cognitivo possível por parte dos usuários, tornando simples a execução de inúmeras tarefas do cotidiano como o envio de mensagens de texto, buscas online e organização de agendas, podendo ser facilmente encontrados em grandes Sistemas Operacionais (SO) de smartphones, a exemplo destacam-se o Android do Google e o iOS da Apple (MATOS, OLIVEIRA, 2021).

Outra aplicação de PLN é o *chatterbot*, um software cujo objetivo é responder perguntas de tal forma que a pessoa que estiver comunicando-se com ele tenha a impressão de estar conversando com outra pessoa. Ou seja, ele emula uma conversa, uma comunicação humana (COMARELLA, CAFÉ, 2008). *Chatterbot* é um *software* de comunicação que funciona dentro de aplicativos de mensagens como *WhatsApp* e *Facebook Messenger*. Por meio deles, os usuários trocam mensagens com empresas, estabelecendo uma conversa com um robô, ou se preferir, com uma máquina, geralmente dotada de Inteligência Artificial.

De acordo com Comarella, Café (2008, p. 59), “as duas características principais que um *chatterbot* deve possuir são a autonomia e a capacidade de interagir com o ambiente”.

Ainda conforme as autoras supracitadas, os *chatterbot* podem ser classificados conforme o seu propósito, como: de entretenimento – procuram divertir o usuário, geralmente simulando “vida artificial”; educacionais – colaboram no desenvolvimento intelectual e no aprendizado do aluno; comerciais – assumem o papel do ser humano em atividades tais como as de suporte ao consumidor, realizam marketing na web, entre outros; e acadêmicos – *chatterbo*t desenvolvidos exclusivamente para estudar as técnicas de desenvolvimento (COMARELLA, CAFÉ, 2008).

Outra aplicação de PLN muito utilizada no nosso cotidiano é o mecanismo de busca que usa o PLN para mostrar resultados relevantes com base em comportamentos de pesquisa semelhantes ou na intenção do usuário, para que pessoas comuns encontrem o que precisam sem precisar saber os termos de pesquisa exatos que devem usar.

Segundo Gabriel (2012, p. 36), mecanismo de busca que são sistemas de recuperação de informações cuja o objetivo está em “auxiliar na busca de informações armazenadas em ambientes computacionais e cuja utilidade pode ser mensurada na relevância (qualidade) e na rapidez de seus resultados (velocidade)”.

Na área de negócio o PLN está sendo empregado pela análise de sentimento ou pesquisas de opinião são abordagens que consistem em identificar um estudo de opiniões, sentimentos, emoções e atitudes, onde é possível detectar, extrair, classificar opiniões, sentimentos e atitudes sobre diversos assuntos. É especialmente útil para identificar tendências da opinião pública nas mídias sociais, para fins de marketing (AMARAL, SILVA, ALMEIDA, 2017).

O PLN tem grande sucesso de aplicação no direito, em que o armazenamento e recuperação de textos jurídicos é um grande desafio devido ao excesso de documentos que são diariamente gerados. Através da técnica de sumarização de texto, cujo objetivo é identificar segmentos relevantes do texto que devem ser incluídos no sumário (extrato) final (SANTOS, Â, 2012; LADEIRA, 2010). Muitas vezes usados para fornecer resumos a partir de depoimentos e de relatórios policiais, ambos geralmente escritos em texto livre (CARRILHO JÚNIOR, 2007).

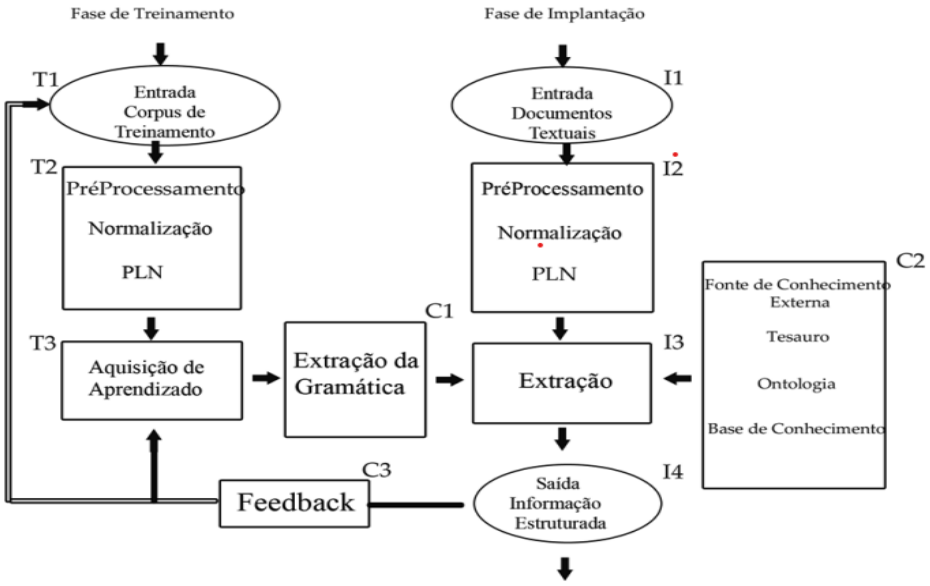
A Sumarização de texto é o processo de destilação das informações mais importantes de um texto, pois visa produzir uma versão resumida do referido texto (JURAFSKY, MARTIN, 2020). Assim, a sumarização de texto pretende criar um documento, a partir de uma ou mais fontes textuais, menor em tamanho, mas mantém algumas ou a maioria das informações contidas nas fontes originais.

Na área da saúde, o PLN tem potencial para contribuir com soluções no âmbito da extração e estruturação de informação clínicas textuais para disponibilizar dados clínicos para uso na tomada de decisão (SOUZA, FELIPE, 2021).

A de extração de informação é uma importante aplicação no campo do PLN e da Linguística, ao extrair automaticamente informações de fontes de informações estruturadas, semiestruturadas ou não estruturadas. Neste caso, a parte do documento que não é relevante pode ser ignorada. Geralmente o problema é abordado no contexto de coleções específicas. Uma abordagem para a técnica é varrer o texto, buscando palavras-chave e extrair dos contextos onde ocorrem tais palavras a informação necessária (CARDOSO, 2004).

A Figura 13 mostra que a arquitetura de um sistema de Extração de Informações normalmente tem duas fases distintas: A fase de treinamento e a fase de implantação, onde T refere-se aos componentes da fase de treinamento, I aos componentes da fase de implantação e C aos componentes de conhecimento.

Figura 13 - Arquitetura básica de um sistema de Extração de Informação



Fonte: Neves, Corrêa, Cavalcanti (2013, p. 38)

Ainda dentro do âmbito da área da saúde é aplicado o reconhecimento de entidades nomeadas, sendo uma subárea de estudo no campo de extração de informação, cuja tarefa é identificar entidades nomeadas, bem como classificá-las dentro de um conjunto de categorias pré-definidas, tais como Pessoa, Organização, Localização, as quais remetem a um referente específico (AMARAL, 2013). Já as abordagens para Reconhecimento de Entidades Nomeadas “tradicionalmente fazem uso de muitas técnicas vindas da Linguística, tais como: etiquetas sintáticas, lema das palavras, prefixos e sufixos, entre outras, para extrair as informações presentes nos textos” (PRIVATTO, 2020, p. 14).

Segundo a dissertação de Garcia (2021), a sua pesquisa teve como objetivo geral desenvolver uma solução que se permitia a identificação de alterações nas frequências dos eventos adversos e queixas técnicas de dispositivos médicos no Brasil em textos livres, por meio da extração de informações, mais especificamente, da aplicação de técnica de Reconhecimento de Entidades Nomeadas.

E na área de educação museal, teve a utilização da tecnologia Watson da IBM na Pinacoteca de São Paulo. Segundo Chiovatto (2019, p. 219), “Watson é um recurso interativo que utiliza IA a fim de simular diálogos com os visitantes”. A autora aponta que a diferença fundamental dos sistemas de IA antigos e o *Deep Learning* é o reconhecimento da fala natural humana e sua interpretação, portanto, possibilidade de mimeses de interação.

A partir da pesquisa realizada é possível categorizar as principais aplicações da área de PLN em quatro grandes categorias e uma categoria de fundamentação de conceitos sobre a temática de PLN.

1 Fundamentação e conceituais: se refere às pesquisas de dados filosóficos, estudo de conceitos e levantamentos de dados dos assuntos selecionados, que são identificados por reflexões, bem como explicações seja na área da CI ou em outros ambientes institucionais, como hospitais, laboratórios dentre outros.

2 Pré-processamento de texto: Essa categoria envolve técnicas que são aplicadas antes da análise propriamente dita do texto. Isso inclui etapas como tokenização (divisão do texto em unidades menores, como palavras ou caracteres), remoção de *stopwords* (palavras comuns que não contribuem para a compreensão do texto, como "o", "e", "um"), normalização de texto (padronização de maiúsculas/minúsculas, remoção de pontuação) e lematização (redução das palavras ao seu lema base). Essa categoria abrange técnicas que envolvem a análise da estrutura gramatical do texto. Isso inclui a identificação de palavras-chave, a determinação da função gramatical de cada palavra (como sujeito, verbo, objeto) e a análise das relações sintáticas entre as palavras. Em resultado dessas técnicas temos a sumarização automática de texto que surge como uma possível solução para reduzir o tempo dos usuários em identificar as informações mais relevantes de um conjunto de documentos textuais. A sumarização automática de texto pode ser definida como a tarefa de geração automática de uma versão condensada (resumo), a partir de um único documento (mono documento) ou de uma coleção de documentos relacionados (multidocumento), mantendo somente as informações mais relevantes.

3 Análise de Semântica e Representação de Texto. Análise de semântica e representação de texto é uma área fundamental no PLN e abrange técnicas que visam compreender o significado dos textos e estabelecer relações semânticas entre eles. A técnica de similaridade e relacionamento semântico se concentra especificamente nesse aspecto, ao medir a proximidade semântica entre termos e avaliar a relação semântica entre elementos textuais.

4 Extração de informações e Mineração de texto: Nessa categoria, as técnicas visam extrair informações específicas e estruturadas do texto. Isso pode envolver a identificação de entidades nomeadas (como nomes de pessoas, organizações, datas), a detecção de relações entre entidades (como "X é o autor de Y") e a extração de informações relevantes de documentos (como resumos

automáticos, categorização de documentos). A Extração de Informações tem objetivo de localizar, estruturar e armazenar a informação relevante de um documento ou de um conjunto de documentos, a fim de propiciar uma futura descoberta de relacionamentos interessantes entre as informações extraídas. Já a Mineração de textos, surge com a finalidade de resolver problemas de descoberta de conhecimento em bases de texto, oferecendo um conjunto de métodos que permite a navegação, organização e descoberta inteligente de informação em bases de dados não estruturadas.

5 Modelagem de tópicos e Classificação de texto: Essa categoria engloba técnicas que visam identificar tópicos ou temas presentes nos textos. Isso inclui a modelagem de tópicos usando métodos como a Alocação Latente de Dirichlet (LDA), onde são inferidos os tópicos subjacentes a um conjunto de documentos. Cada documento consiste em várias palavras e cada tópico pode ser associado a algumas palavras. O objetivo do LDA é encontrar os tópicos aos quais o documento pertence, com base nas palavras nele contidas. Ele pressupõe que documentos com tópicos semelhantes usarão um grupo de palavras semelhantes. Isso permite que os documentos mapeiem a distribuição de probabilidade sobre tópicos latentes e tópicos são distribuição de probabilidade. Além disso, a classificação de texto é usada para atribuir rótulos ou categorias a um texto com base em seu conteúdo, como classificar um e-mail como spam ou categorizar documentos em determinadas áreas temáticas.

4 ANÁLISE E DISCUSSÃO

Após a primeira etapa da pesquisa, que se concentrou em construir um referencial teórico para um melhor entendimento a respeito da área da Ciência da Informação e o campo de Processamento de Linguagem Natural, foi realizada uma revisão bibliográfica com o objetivo de compreender essa temática.

A segunda etapa, com base no livro de Bardin (2011), se desdobrou em 3 fases:

Na primeira fase foi feita a seleção do material nas Bases de Dados BRAPCI e SCOPUS.

Na segunda fase, os artigos selecionados foram divididos em categorias, considerando as quatro principais categorias de aplicações na área de PLN, apresentadas no capítulo anterior e os artigos com base em fundamentação e conceituação. As categorias criadas foram: Artigos de Fundamentação e Conceituais, Pré-processamento de texto; Análise de Semântica e Representação de Texto; Extração de informações e Mineração de texto; Modelagem de tópicos e Classificação de texto.

O Quadro 3 abaixo apresenta a quantidade de artigos em cada categoria elencada.

Quadro 3 - Quantidade de artigos analisados

Categorias	Quantidades de artigos
Artigos de Fundamentação e Conceituais	9
Artigos com foco na categoria de Pré-processamento de texto	5
Artigos com foco na categoria de Análise semântica e Representação de texto	9
Artigos com foco na categoria Extração de informações e Mineração de textos	34
Artigos com foco na categoria de Modelagem de tópicos e Classificação de texto	11

Fonte: Autoria própria, 2023

O subcapítulo seguinte corresponde à terceira fase, onde foi apresentada a análise e discussão dos resultados. No próximo tópico, apresenta a categorização dos

artigos selecionados com os temas Processamento de Linguagem Natural e Ciência da Informação.

4.1 Categorização dos Artigos Selecionados

Através da leitura dos artigos e após verificar os objetivos de cada texto, foi possível iniciar a categorização.

A partir da análise realizada, foram elencadas cinco categorias, nas quais quatro categorias foram definidas pelas principais aplicações de PLN e uma por artigos conceituais. As categorias foram definidas da seguinte maneira: Artigos de Fundamentação e Conceituais, Pré-processamento de texto; Análise de Semântica e Representação de Texto; Extração de informações e Mineração de texto; Modelagem de tópicos e Classificação de texto.

Na análise, foi feita uma divisão que busca detalhar informações relevantes e servem para categorizar os artigos escolhidos. Portanto, no próximo tópico, foram apresentadas as análises dos artigos classificados de acordo com as categorias elaboradas. Em cada categoria, apresentamos um quadro resumido dos artigos analisados, com informações de autoria, título e ano. Os Apêndices A, B, C, D, E traz mais dados desses artigos.

4.1.1 Categoria de artigos de Fundamentação e conceituais

Nessa categoria foram apresentados como os autores dos artigos analisados definiram a área de Processamento de Linguagem Natural, com textos conceituais e explicativos, mostrando o interesse crescente dos pesquisadores em assuntos relacionados ao PLN não somente entre profissionais da Computação, mas em várias áreas, pois a aplicação de PLN, gera interesse em várias áreas de atuação.

Quadro 4 - Artigos selecionados na categoria de Fundamentação e conceituais

TÍTULOS	AUTORES	ANO
Using natural language processing techniques to inform research on nanotechnology	LEWINSKI, N. A.; MCINNES, B. T.	2015
Um estudo bibliográfico sobre ligação de entidades	MAIA, E. H. B.; BAX, M. P.	2016

Análise da produção científica do periódico JASIS&T sob a ótica dos três paradigmas da Ciência da Informação	ALVES, R. P. S.; CURTY, R. G.; TREVISAN, G. L.	2018
Estudo sobre contribuição da Ciência da Informação em pesquisas sobre Tecnologias Assistivas	SOUZA, O.; TABOSA, H. R.	2018
Uso de ferramentas computacionais como auxílio ao método de mapeamento cruzado entre terminologias clínicas	GOMES, D. C. *et al.*	2019
Metodologias, ferramentas e aplicações da inteligência artificial nas diferentes linhas do combate a Covid-19	NEVES, B. C.	2020
Sistemas de Indexação automática por atribuição: uma análise comparativa	SILVA, S. R B. CORRÊA, R. F.	2020
Absorção das Tarefas de Processamento de Linguagem Natural (NLP) pela Ciência da Informação (CI): uma revisão da literatura para tangibilização do uso de NLP pela CI.	FALCÃO, L. C. J.; LOPES, B.; SOUZA, R. R.	2021
O Processamento de Linguagem Natural nos Estudos Métricos da Informação: uma análise dos artigos indexados pela Web of Science (2000-2019)	PUERTA-DíAZ, M. *et al.*	2021

Fonte: Autoria própria, 2023

Dos artigos analisados, 9 foram classificados nesta categoria, sendo 7 localizados a partir da Base de Dados BRAPCI e 2 a partir da Base de Dados SCOPUS. Abaixo, apresentamos uma análise sucinta de cada artigo selecionado:

Lewinsky, McInnes (2015) buscaram revisar os diferentes métodos de informática que têm sido aplicados à mineração de patentes, caracterização de nanomateriais/dispositivos, nanomedicina e avaliação de risco ambiental. Para isso, eles focaram nas ferramentas que utilizam Processamento de Linguagem Natural. Em sua revisão, os autores conceituaram que o PLN “envolve o uso de computadores para realizar tarefas práticas envolvendo a linguagem escrita, como extrair e analisar informações de um texto não estruturado” (LEWINSKI, MCINNES, 2015, p. 1440). Também apontaram que o que separa as aplicações de PLN de outros sistemas de processamento de dados é o uso do conhecimento sobre a linguagem humana. Como mencionado anteriormente, segundo o autor Nascimento (2011, p. 12), “a linguística é a ciência que estuda toda linguagem verbal ou escrita que faz parte da língua, tendo nela sua matéria de estudo e reflexão”.

Maia, Bax (2016) apresentaram quais os problemas relacionados à Ligação de Entidades (LE), suas aplicações típicas, bem como sintetizar suas principais

abordagens no contexto da ligação de conceitos. Em sua pesquisa os autores explicaram que PLN “estuda os problemas da geração e compreensão automática da informação armazenada em linguagem natural. A informação pode ser encontrada em bases de dados estruturadas ou não estruturadas” (MAIA, BAX, 2016, p. 246). Para extrair as informações dessas bases de dados não é uma tarefa simples, e uma das tarefas que podem auxiliar é a Ligação de Entidades.

Alves, Curty, Trevisan (2018) apontaram que o termo de Processamento de Linguagem Natural tem o seu aparecimento no *corpus* de análise de sua pesquisa no segundo tempo (Relação entre Informação e Conhecimento de Barreto), porém a pesquisa também indicou que o termo foi mais aplicado em 1995. Que esse dado pode ser sustentado pelo autor Chowdhury (2003) que ao descrever o PLN ressalta-se como uma área de pesquisa de interesse mais recente na CI, que a inter-relação com a Ciência da Computação e a Linguística, visa “aplicar técnicas e ferramentas para a extração automática e análise de textos e de linguagem falada, e que teve maior proeminência a partir da proliferação da *web,* das bibliotecas digitais e dos avanços em Inteligência Artificial” (ALVES, CURTY, TREVISAN, 2018, p. 19). A pesquisa dos autores supracitados demonstrou o advento do termo de PLN no segundo tempo paradigmático proposto por Barreto (2007) na CI. Essa pesquisa confirma a fala de Saracevic (1996) que a CI é campo inexoravelmente interdisciplinar, de interlocuções constantes e mutáveis a partir de refinamentos com diferentes disciplinas.

Souza, Tabosa (2018) procuraram conhecer as aplicações e as limitações dos estudos sobre acessibilidade informacional desenvolvido na CI. Os autores argumentaram em sua pesquisa que uma aplicação direta do PLN que contribuiria na assistividade informacional seria a produção automática de resumos de textos, porém com mínimo possível de perda semântica e também sintetizar voz no idioma dos leitores.

Gomes *et al.* (2019) em sua pesquisa, tiveram como objetivo, refletir sobre o uso de ferramentas computacionais no método de mapeamento cruzado entre terminologias clínicas. Para isso, destacaram o uso da tarefa de Extração automática de termos, a qual envolve a resolução de termos simples e compostos e pode ser baseada em estatística, linguística e/ou conhecimento. Para os autores, o PLN pode utilizar análise morfológica, sintática, semântica e pragmática. Como citado anteriormente neste trabalho, a Linguística tem vertentes que correspondem a

diferentes tipos de análise dos diferentes aspectos da língua. Nesse momento podemos observar a contribuição da Linguística e da Terminologia para o Processamento de Linguagem Natural. Pois segundo Pontes (1997, p. 47), “o termo, que é o objeto de estudo da Terminologia, é basicamente um signo linguístico formado de uma denominação (significante) e um conceito (significado)”. Nisso podemos concluir a inter-relação da Linguística, Terminologia com o Processamento de Linguagem Natural e pôr fim a sua relação com a Ciência da Informação, que busca a padronização dos termos.

A autora Neves (2020) apresentou em sua pesquisa um conjunto de metodologias, ferramentas e aplicações da Inteligência Artificial nas diferentes linhas do combate ao novo coronavírus e a Covid-19. E uma dessas ferramentas é o “Processamento de Linguagem Natural que se refere, especificamente, a capacidade para que as máquinas possam compreender a linguagem humana e a partir disso possam extrair de textos a informação (significado) contida neles” (NEVES, 2020, p. 48).

Silva, Correa (2020) realizaram uma análise comparativa de dois sistemas de indexação automática por atribuição multilíngue: SISA e MAUI. Nessa análise os autores puderam identificar que nas características dos sistemas SISA e MAUI que algumas operações são aplicações de PLN, no qual tem o objetivo de melhorar a qualidade do processamento das informações via extração de termos relevantes.

Falcão, Lopes, Souza (2021) procuraram identificar quais recursos em PLN vêm sendo desenvolvidos e em quais áreas eles têm sido mais aplicados. E apontaram o quanto a Ciência da Informação tem utilizado esses recursos em suas atividades. Os autores definiram Processamento de Linguagem Natural, enquanto disciplina é compreendida como um elo entre Ciência da Computação e Ciência da Informação. E que o PLN também conhecido como linguística computacional, pertence ao campo da Ciência da Computação e da Linguística como um subcampo da Inteligência Artificial.

Os autores Puerta-Díaz *et al.* (2021) buscaram em sua pesquisa identificar a estrutura científica internacional das pesquisas que vinculam o uso do Processamento de Linguagem Natural no campo dos estudos métricos da informação. Para isso, os autores definiram que “PLN é uma área de pesquisa e aplicação que explora como os computadores podem ser usados para entender e manipular texto ou fala em linguagem natural para fazer coisas úteis” (PUERTA-DÍAZ *et al.*, 2021, p. 4). Ainda

em sua pesquisa, os autores mencionaram que o objetivo do PLN é realizar o processamento de linguagem idêntico ao feito pelos humanos.

4.1.2 Categoria de Pré-processamento de texto

Nessa categoria são apresentados como os autores realizaram a aplicação de Pré-processamento de texto em suas pesquisas.

Quadro 5 - Artigos selecionados na categoria de Pré-processamento de texto

TÍTULOS	AUTORES	ANO
Query Snowball: A Co-occurrence-based Approach to Multi-document Summarization for Question Answering	MORITA, H.; SAKAI, T.; OKUMURA, M.	2011
An automated knowledge-based textual summarization system for longitudinal, multivariate clinical data	GOLDSTEIN, A.; SHAHAR, Y.	2016
The UAB Informatics Institute and 2016 CEGS N-GRID De-Identification Shared Task Challenge	BUI, D. D. A.; WYATT, M.; CIMINO, J. J.	2018
Base de Normas Jurídicas Brasileiras: uma iniciativa de Open Government Data	MARTIM, H.; LIMA, J. A. O.; ARAUJO, L. C.	2018
Avaliação do desempenho de um software de sumarização automática de textos	TABOSA, H. R. *et al*.	2020

Fonte: Autoria própria, 2023

Dos artigos analisados, 5 foram classificados nesta categoria, sendo 2 localizados a partir da Base de Dados BRAPCI e 3 a partir da Base de Dados SCOPUS. Abaixo, apresentamos uma análise sucinta de cada artigo selecionado:

Os autores Morita, Sakai, Okumura (2011), procuraram desenvolver algoritmos para codificar texto clínico em representações que possam ser usadas para uma variedade de tarefas de fenotipagem. Segundo os autores, a sumarização automática de textos visa diminuir a quantidade de texto que o usuário tem que ler enquanto preserva conteúdos importantes. Conforme os autores argumentaram sobre o cenário atual de sobrecarga informacional, existe um constante interesse no desenvolvimento de ferramentas capazes de recuperar, classificar e sintetizar informações relevantes de maneira rápida e eficiente, por esse contexto, podemos ressaltar a importância da ferramenta de sumarização automática de textos.

Goldstein, Shahar (2016) desenvolveram um sistema inteligente de sumarização de texto livre para dados clínicos longitudinais e multivariados. O sistema

tem o objetivo de estruturar o documento e organizar os dados brutos e abstratos para melhorar a qualidade de atendimento dos pacientes.

Bui, Wyatt, Cimino (2018), buscaram em seu trabalho a desidentificação automatizada ou semiautomática de dados clínicos. Para a desidentificação automática, na fase do pré-processamento (PLN) foi utilizado a ferramenta Stanford NLP. Também foi utilizado o termo do PHI para classificação. O artigo trabalha sobre dados sensíveis médicos.

Martim, Lima, Araújo (2018) propuseram apresentar uma série de transformações automática aplicada ao arcabouço de leis federais de modo a estruturar a informação descrita nesses documentos, com o propósito de prepará-las para diferentes tipos de interpretações automáticas, como identificação de entidades nomeadas, definições, remissões, eventos de criação, alteração e encerramento de instituições jurídicas, recuperação da versão vigente de uma lei no tempo, com intuito de contribuir com cientistas da informação e pesquisadores em geral. Martim, Lima, Araújo (2018) afirmaram que o PLN tem evoluído rapidamente nos últimos anos, muito pela aplicação de técnicas de Inteligência Artificial.

Tabosa *et a*l. (2020) buscaram avaliar o software (protótipo) desenvolvido por eles em seu trabalho anterior. Um software que seria capaz de elaborar resumos automáticos de textos baseado em técnicas de PLN e estatísticas de frequência de palavras. Os autores descreveram que no processo de sumarização automática de textos é necessária uma etapa semelhante ao processo de indexação. E os autores reforçaram que as técnicas de indexação automática por extração e por atribuição podem ser combinadas para a produção de resumos de textos. Os autores ainda destacaram que algumas fases de PLN são de interesse para o desenvolvimento de um sumarizador de textos: morfologia, sintaxe, semântica e pragmática. Ressaltando que essas fases já foram apresentadas nesse trabalho em tipos de análises linguísticas.

4.1.3 Categoria de Análise Semântica e Representação de texto

Nessa categoria são apresentados como os autores realizaram as aplicações de Análise Semântica e Representação de texto em suas pesquisas.

Quadro 6 - Artigos selecionados na categoria Análise Semântica e Representação de texto

TÍTULOS	AUTORES	ANO
Discerning truth from deception: Human judgments & automation efforts	RUBIN, V.; CONROY, N.	2012
Information manipulation classification theory for LIS and NLP	RUBIN, V. L.; CHEN, Y.	2012
An Evaluative Baseline for Geo-Semantic Relatedness and Similarity	BALLATORE, A.; BERTOLOTTO, M.; WILSON, D. C.	2014
A framework for unifying ontology-based semantic similarity measures: A study in the biomedical domai	HARISPE, S. *et al.*	2014
20 Deception Detection for News: Three Types of Fakes	RUBIN, V. L.; CHEN, Y.; CONROY, N. K.	2015
A Novel Approach for Computing Semantic Relatedness of Geographic Terms	SOLEIMANDARABI, M. N.; MIRROSHANDEL, S. A.	2015
CoTO: A novel approach for fuzzy aggregation of semantic similarity measures	MARTINEZ-GIL, J.	2016
Redes complexas de homônimos para análise semântica textual	SANTOS, J. *et al.*	2017
Inteligência artificial e ferramentas da Web Semântica aplicadas a recuperação da informação: um modelo conceitual com foco na linguagem natural	CONEGLIAN, C. S.; SANTARÉM SEGUNDO, J. E.	2022

Fonte: Autoria própria, 2023

Dos artigos analisados, 9 foram classificados nesta categoria, sendo 2 localizados a partir da Base de Dados BRAPCI e 7 a partir da Base de Dados SCOPUS. Abaixo, apresentamos uma análise sucinta de cada artigo selecionado:

Rubin, Conroy (2012) tiveram como objetivo na sua pesquisa em andamento, informar a criação de ferramentas para detecção de engano. As autoras utilizaram algoritmos de aprendizado de máquina e extração de recursos linguísticos para criação das ferramentas. Rubin, Conroy (2012) apontaram que com o crescente uso da Comunicação Mediada por Computador (CMC) em todos os aspectos da civilização moderna, a detecção de enganos em CMC emergiu como uma questão importante nas comunicações cotidianas e agora é de interesse no amplo campo da Biblioteconomia e Ciência da Informação. Ainda afirmaram que a detecção automatizada de enganos é uma contribuição recentemente acessível do PLN e aprendizado de máquina.

Em continuidade, as autoras Rubin, Chen (2012), utilizaram da técnica de PLN, utilizando um conjunto pré-definidos de dimensões (facetas) características

mutuamente exclusivas em cada *continuum* (focos). As autoras utilizaram a classificação facetada como conjunto de pré-requisito na técnica de PLN. Rubin, Chen (2012) propuseram em sua pesquisa, uma contribuição na área de Biblioteconomia e Ciência da Informação, com uma compreensão diferenciada das variedades de manipulação da informação. Para o PLN cria um potencial para reconhecimento automatizado e adaptabilidade da tecnologia de detecção de engano para a identificação de outras variedades de manipulação de informações com base em semelhanças.

Ballatore, Bertolotto, Wilson (2014) em seu artigo, tiveram como objetivo, debater uma noção de relação geosemântica baseada nos campos semânticos de Lehrer e a comparar com similaridade geosemântica. Os autores citaram que “a relação envolve todas as relações semânticas, incluindo sinonímia, antonímia, hiponímia, hipernímia, holonímia, meronímia, causalidade, contiguidade temporal, função, proximidade e contenção” (BALLATORE, BERTOLOTTO, WILSON, 2014, p. 6). E para obter similaridade semântica, a observação deve ser restrita à taxonomia, é uma relação entre os termos.

Os autores Harispe *et al.* (2014) analisaram as medidas existentes baseadas na ontologia para identificar os elementos centrais da avaliação da similaridade semântica. Segundo os autores (2014, p. 1) “medidas de similaridade semântica são utilizadas para estimar a similaridade de conceitos definidos em ontologias e, assim, avaliar a proximidade semântica dos recursos por elas indexados”. Para esse feito as medidas de similaridade em ontologia visam estimar quão semelhantes são os significados dos conceitos de acordo com as evidências taxonômicas modeladas na ontologia. Elas são utilizadas em uma extensa gama de aplicações: para recuperação da informação na literatura científica, para desambiguar textos, para o processamento automático de mensagens de texto, para os sistemas de diálogo de saúde, para a consulta de linguagem natural de bancos de dados e para as respostas a perguntas.

Rubin, Chen, Conroy (2015) tiveram como objetivo principal em sua pesquisa levantar os requisitos de corpora de notícias falsas (para adequação em análise textual e modelagem preditiva). Rubin, Chen, Conroy (2015, p.1), citaram que “filtrar, vetar e verificar informações on-line continuam a ser essenciais em Biblioteconomia e Ciência da Informação, pois as linhas entre notícias tradicionais e informações on-line estão se confundindo”.

Os autores Soleimandarabi, Mirroshandel (2015) apresentaram uma medida

baseada em *corpus* para computar a relação semântica de termos geográficos com base em sua definição lexical extraída do léxico geográfico e empregando a Wikipédia como conhecimento de fundo. Para esses autores a relação semântica computacional é manifestada como uma tarefa fundamental no PLN e seu objetivo é identificar uma medida que possa expressar a força da associação semântica entre um par de conceitos incluindo relações clássicas e não clássicas. A relação semântica geográfica, tem ampla aplicação na descoberta de termos em mapa geográfico, recuperação de geoinformação e Ciência da Informação Geográfica.

O autor Martinez-Gil (2016) apresentou o CoTO (*Consensus* ou *Trade-Off*) como uma solução baseada em lógica *fuzzy* para agregação de valores de similaridade semântica. Em seu artigo o autor apontou a importância da medição precisa da semelhança semântica para o campo da computação, como um método chave para várias tarefas, como: o agrupamento de dados para detectar e agrupar os assuntos mais semelhantes, a correspondência de dados que consiste em encontrar alguns dados que se referem ao mesmo conceito em diferentes fontes de dados, a mineração de dados onde o uso de medidas de similaridade semântica apropriadas pode ajudar a facilitar tanto os processos de classificação de texto quanto a descoberta de padrões em textos grandes ou tradução automática onde a detecção de pares de termos expressos em diferentes línguas.

Santos *et al.* (2017) apresentaram o processo de utilização de redes complexas como base de dados comparativa para determinar, através do contexto, o significado de palavras que expressam posicionamentos distintos. Além disso, são classificados com mesma morfologia e sintaxe, como ocorre com alguns homônimos. Santos *et al.* (2017) afirmaram que PLN é uma área que estuda problemas de geração e compreensão automática de linguagens humana entre homem e máquina. “Para analisar um texto através de PLN é preciso considerar aspectos morfológicos, sintáticos e semânticos, de modo que permita o tratamento do texto e o seu entendimento” (SANTOS *et al*. 2017, p. 295).

Coneglian, Santarém Segundo (2022) tiveram como objetivo na sua pesquisa propor um modelo conceitual de recuperação da informação, a partir da aproximação da linguagem computacional com a linguagem natural, utilizando os princípios da representação da informação, para que o significado e o contexto dos dados estejam explícitos para o processo da busca. Coneglian, Santarém Segundo (2022) afirmaram em sua pesquisa, que fundamenta as relações interdisciplinares entre Ciência da

Informação e Ciência da Computação, após a proposta de um novo modelo de recuperação da informação, pautado na Inteligência Artificial e seus diversos campos, e a Web Semântica. Os autores utilizaram o modelo de Question Answering, seja modelo QA.

4.1.4 Categoria de Extração de Informações e Mineração de texto

Nessa categoria são apresentados como os autores realizaram as aplicações de Extração de Informações e Mineração de texto em suas pesquisas.

Quadro 7 - Artigos selecionados na categoria aplicação de Extração de Informações e Mineração de texto

TÍTULOS	AUTORES	ANO
Descoberta de Conhecimento em Texto aplicada a um Sistema de Atendimento ao Consumidor	SCHIESSL, M.; BRÄSCHER, M.	2011
Creating Chinese-English comparable corpora	HUANG, D.; WANG, S.; REN, F.	2013
Assessing the role of a medication-indication resource in the treatment relation extraction from clinical text	BEJAN, C. A.; WEI, W-Q.; DENNY, J. C.	2014
Automatic abstraction of imaging observations with their characteristics from mammography reports	BOZKURT, S.; *et al.*	2014
Utility-preserving privacy protection of textual healthcare documents	SÁNCHEZ, D.; BATET, M.; VIEJO, A.	2014
Fundamentos em processamento de linguagem natural: uma proposta para extração de bigramas	SILVA, E. M.; SOUZA, R. R.	2014
Text Mining and Big Data Analytics for Retrospective Analysis of Clinical Texts from Outpatient Care	BOYTCHEVA, S. *et al.*	2015
A Global Optimization Approach to Multi-Polarity Sentiment Analysis	LI, X.; LI, J.; WU, Y.	2015
Extracting Development Tasks to Navigate Software Documentation	TREUDE, C.; ROBILLARD, M. P.; DAGENAIS, B. E.	2015
Multilayered temporal modeling for the clinical domain	LIN, C. *et al.*	2016
A novel web informatics approach for automated surveillance of cancer mortality trends	TOURASSIA, G.; YOONA, H-J.; XUB, S.	2016
EliIE: An open-source information extraction system for clinical trial eligibility criteria	KANG, T. *et al.*	2017
Electronic Health Record Phenotypes for Precision Medicine: Perspectives and Caveats from Treatment of Breast Cancer at a Single Institution	BREITENSTEIN, M. K. *et al.*	2018
The UK Online Gender Audit 2018: A comprehensive audit of gender within the UK's online environment	HULUBA, A-M.; KINGDON, J.; MCLAREN, I.	2018
CogStack - experiences of deploying integrated information retrieval and extraction services in a large	JACKSON, R. *et al.*	2018

National Health Service Foundation Trust hospital		
Developing a healthcare dataset information resource (DIR) based on Semantic Web	SHI, J. *et al.*	2018
Automatic Detection of Negated Findings in Radiological Reports for Spanish Language: Methodology Based on Lexicon-Grammatical Information Processing	KOZA, W.; *et al.*	2019
Agrupamento automático de notícias de jornais on-line usando técnicas de Machine Learning para clustering de textos no idioma português	MAGALHÃES, L. H.; SOUZA, R. R.	2019
Exemplo de Extração de Definições em Textos Articulados de Normas Jurídicas com o apoio do Processamento de Linguagem Natural	TEIXEIRA, W. R.; e*t al.*	2019
Natural Language Processing Combined with ICD-9-CM Codes as a Novel Method to Study the Epidemiology of Allergic Drug Reactions	BANERJI, A. *et al.*	2020
Exploring the Potential of Twitter to Understand Traffic Events and Their Locations in Greater Mumbai, India	DAS, R. D.; PURVES, R. S.	2020
University Learning and Anti-Plagiarism Back-End Services	KOLHAR, M.; ALAMEEN, A.	2020
A Comparison of the Development of Medical Informatics in China and That in Western Countries from 2008 to 2018: A Bibliometric Analysis of Official Journal Publications	LIANG, J. *et al.*	2020
Application of entity linking to identify research fronts and trends	MARRONE, M.	2020
Text mining and semantic triples: Spatial analyses of text in applied humanitarian forensic research	MIRANKER, M.; GIORDANO, A.	2020
Impact of different electronic cohort definitions to identify patients with atrial fibrillation from the electronic medical record	SHAH, R. U. *et al.*	2020
Clinical Relation Extraction Using Transformer-based Models	YANG, X. *et al.*	2020
Network graph representation of COVID-19 scientific publications to aid knowledge discovery	CERNILE, G. *et al.*	2021
Measuring the Interactions Between Health Demand, Informatics Supply, and Technological Applications in Digital Medical Innovation for China: Content Mapping and Analysis	DU, J.; CHEN, T.; ZHANG, L.	2021
Feasibility of capturing real-world data from health information technology systems at multiple centers to assess cardiac ablation device outcomes: A fit-for-purpose informatics analysis report	JIANG, G.; *et al.*	2021
User Behaviors and User-Generated Content in Chinese Online Health Communities: Comparative Study	LEI, Y.; XU, S.; ZHOU, L.	2021
O fluxo temporal de termos relevantes: uma análise em teses da UFMG de 2007 a 2018 nas ciências humanas	MESQUITA, L. A. L.; DIAS, C. C.; SOUZA, R. R.	2021
Processamento de Linguagem Natural aplicado à anamneses do domínio da ginecologia	SOUZA, A. D.; FELIPE, E. R.	2021

Using parsed and annotated corpora to analyze parliamentarians' talk in Finland	ANDRUSHCHENKO, M. *et al.*	2022

Fonte: Autoria própria, 2023

Dos artigos analisados, 34 foram classificados nesta categoria, sendo 6 localizados a partir da Base de Dados BRAPCI e 28 a partir da Base de Dados SCOPUS. Abaixo, apresentamos uma análise sucinta de cada artigo selecionado:

Segundo os autores Schiessl, Bräscher (2011), o objetivo da sua pesquisa é aplicar a Descoberta de Conhecimento em Texto na base do SAC de uma instituição bancária, com a finalidade de criar automaticamente agrupamentos de documentos para posterior criação de um modelo categorizador automático dos novos documentos recebidos diariamente, para essa tarefa os autores aplicaram a técnica de mineração de textos. Os autores afirmaram que a integração de técnicas de PLN e Descoberta de Conhecimento em Texto, tem o objetivo de automatizar o processo de transformação de dados textuais em informação para possibilitar a aquisição do conhecimento e nesse sentido, vem contribuir enormemente com a Ciência da Informação no que tange ao tratamento e recuperação da informação.

Huang, Wang, Ren (2013) propuseram em sua pesquisa a criação de corpora comparáveis Chinês-Inglês. Huang, Wang, Ren (2013, p. 1853) apontaram que corpora multilíngues, paralelos ou comparáveis, são recursos valiosos para recuperação de informações em vários idiomas e muitos outros estudos linguísticos, como Análise de Discurso, Análise Pragmática, Extração de Terminologia e Engenharia do Conhecimento.

Bejan, Wei, Denny, (2014) tiveram como o principal objetivo avaliar vários métodos de extração de relações de tratamento de notas clínicas. Segundo os autores, o reconhecimento de relações terapêuticas entre conceitos médicos descritos em documentação clínica é uma tarefa desafiadora para aplicações atuais de PLN.

Bozkurt *et al.* (2014) buscaram desenvolver métodos de PLN para reconhecer cada lesão em relatórios de mamografia de texto livre e extrair suas relações correspondentes, produzindo um quadro de informações completo para cada lesão. Para isso, construíram um *pipeline* de extração de informações de PLN no kit de ferramentas de PLN da General Architecture for Text Engineering (GATE). Os autores indicaram que há vários sistemas de PLN desenvolvidos para extração de informações de registros clínicos. Em geral, esses sistemas anotam estruturas sintáticas, realizam reconhecimento de entidade nomeada, mapeiam extensões de

texto para conceitos de um vocabulário ou ontologia controlada e identificam o contexto de negação de entidades nomeadas.

Sánchez, Batet, Viejo (2014) utilizaram em seu trabalho a tarefa de detecção automática de sintagmas nominais, para higienizar automaticamente um documento médico textual de acordo com certos critérios de privacidade. Os autores destacaram que utilizaram várias ferramentas de PLN, como o kit de ferramenta de OpenNLP, que fornece suporte para as tarefas de: detecção de sentenças, lematização, análise sintática, extração de entidades nomeadas.

Silva, Souza (2014) investigaram em sua pesquisa, como extrair o significado do texto através de um conjunto de expressões multipalavras identificadas. Segundo Silva, Souza (2014, p. 2) as expressões multipalavras “as tornam relevantes no tratamento dos recursos lexicais, os quais são importantes insumos informacionais para muitas aplicações relacionadas ao PLN, tais como: Recuperação da Informação (RI), tradução automática, sumarização de texto, etc.”

Boytcheva *et al.* (2015) afirmaram que no decorrer da última década foram desenvolvidos vários sistemas de Extração de Informação para análise de textos clínicos – para extração de diagnóstico, identificação de medicamentos e posologia, reconhecimento de queixas e eventos relacionados, fatores de risco, etc.

Os autores Li, LI, Wu (2015) buscaram em sua pesquisa propor uma abordagem de análise de sentimento em otimização global. Para os autores, o rápido desenvolvimento das mídias sociais, a análise de sentimentos tornou-se uma importante técnica de mineração de mídias sociais. Os pesquisadores ressaltaram que “a análise de sentimentos também é conhecida como computação de polaridade emocional, extração de opinião ou classificação semântica” (LI, LI, WU, 2015, p. 2). É definido como o processo de identificar o sentimento e opinião (por exemplo, positivo, negativo ou neutro) em um determinado texto. Podemos evidenciar que a técnica de análise de sentimento pode ser aplicada em várias áreas, como em biblioteca para melhorar o atendimento dos usuários.

Treude, Robillard, Dagenais (2014) desenvolveram uma técnica para extrair automaticamente as tarefas de desenvolvimento da documentação do softwares. Para isso utilizaram as dependências gramaticais entre as palavras, como detectado pelo analisador Stanford NLP.

Lin *et al.* (2016) buscaram desenvolver um sistema de descoberta de relações temporais de código aberto para o domínio clínico. Lin *et al.* (2016), explicaram que a

descoberta automática da relação temporal tem a capacidade de aumentar drasticamente a compreensão de muitos fenômenos médicos, como a progressão da doença, os efeitos longitudinais de medicamentos e o curso clínico de um paciente.

Tourassia, Yoona, Xub (2016) propuseram uma nova abordagem de monitoramento automatizado das tendências de mortalidade por câncer. A abordagem envolve coleta automatizada e mineração de textos de obituários online para derivar a distribuição etária, tendências geoespaciais e temporais das mortes por câncer nos EUA. Em sua pesquisa, os autores destacaram que a mineração de conteúdo online na área da saúde pode ser considerada como uma nova abordagem de aquisição de informações, pode ser particularmente vantajosa para descobertas epidemiológicas que exigem coleta e curadoria de dados demorados.

Kang *et al.* (2017) tiveram como objetivo em seu estudo desenvolver um sistema de extração de informações de código aberto chamado *Elegibility Criteria Information Extração* (EliIE). Os autores afirmaram que, “a extração de informações refere-se à tarefa de extrair automaticamente semântica estruturada (por exemplo, entidades, relações e eventos) a partir de texto não estruturado” (KANG *et al.*, 2017, p. 1063).

Breitenstein *et al.* (2018) afirmaram que as aplicações de PLN baseado em regras têm anotado efetivamente elementos de dados clínicos de câncer de mama de relatórios de patologia supervisionados usando apenas algumas notas de treinamento.

Huluba, Kingdon, McLaren (2018) buscaram medir o nível de representação masculina/feminina nos sites das organizações voltadas para o Reino Unido. E uma das ferramentas utilizadas foi a mineração de textos. Os pesquisadores relataram que “os resultados sugerem que as técnicas de IA podem assumir um novo e importante papel na mineração de fontes de dados, de modo a realizar formas detalhadas de análise social e econômica” (HULUBA, KINGDON, MCLAREN, 2018, p. 3). A pesquisa utilizou a técnica de PLN para “ler” sistematicamente todo o domínio online do Reino Unidos, de modo a coletar informações relacionadas ao gênero. Dentro da CI o trabalho realizou a curadoria dos dados e organizou a informação.

Jackson *et al.* (2018) buscaram criar e implantar uma arquitetura de recuperação e extração de informações estruturadas e não estruturadas de baixo custo no *King's College Hospital*. Autores afirmaram que as tecnologias de recuperação da informação têm o propósito declarado de fornece a capacidade de

filtrar quantidades muito grandes de informações estruturadas e não estruturadas e retornar resultados relevantes em alta velocidade.

Shi *et al.* (2018) desenvolveram métodos para extrações manuais de documentações de conjuntos de dados, bem como extrações semiautomáticas de publicações relacionadas, usando abordagens baseadas em PLN. Para isso desenvolveram um componente de consulta semântica para recuperação de conhecimento.

Koza *et al*. (2019) apresentaram uma metodologia para o reconhecimento automático de achados negados em laudos radiológicos considerando informações morfológicas, sintáticas e semânticas. Os autores elaboraram uma série de regras de processamento de informações lexicais e sintáticas. Isso exigiu o desenvolvimento de um dicionário eletrônico de Terminologia médica e gramática de informática.

Magalhães, Souza (2019) tiveram como objetivo principal do seu trabalho criar aglomerados de notícias a partir de uma amostra coletada dos principais jornais on-line. Para tal propósito, os autores utilizaram a técnica de Clustering. Segundo Magalhães, Souza (2019, p. 1), “clusterização é uma técnica de organizar dados em grupos cujos membros apresentam alguma similaridade”. Essa técnica pode ser a solução automatizada capaz de recuperar e comparar as notícias em destaque na mídia.

Teixeira *et al.* (2019) buscaram auxiliar o gestor da informação jurídica a extrair definições de textos de normas jurídicas para apoiar a elaboração de sistemas de organização do conhecimento, tais como, glossários, tesauros e ontologias.

Banerji *et al.* (2020) tiveram como objetivo principal em sua pesquisa, desenvolver e validar um novo algoritmo de PLN para extrair informações de documentos clínicos de texto livre. Banerji *et al*. (2020, p. 8) citaram que “o campo da PLN está crescendo em um ritmo notavelmente rápido e se tornando maduro o suficiente para que as aplicações práticas tenham um impacto clínico significativo”.

Das, Purves (2020) buscaram extrair todas as localizações de eventos de trânsito em tweets relevantes. O foco principal do trabalho foi a recuperação de informações geográficas pela plataforma do Twitter.

Kolhar, Alameen (2021) em sua pesquisa investigaram um método para detectar plágio aplicando o conceito central de texto, que são associações semânticas de palavras e sua composição sintática. E para esse efeito, foram utilizadas várias ferramentas de PLN.

Liang *et al*. (2020) buscaram apresentar uma comparação entre a informática médica entre a China e os países ocidentais. Nesse estudo eles utilizaram a técnica de clustering e diagramas de rede de coocorrência. Em seu estudo Liang *et al*. (2020, p. 2) explicaram que, “a informática médica pode ser definida como a aquisição, armazenamento, recuperação e uso de informações de saúde” e com a aplicação cada vez mais ampla da Ciência da Computação e da Tecnologia da Informação nas áreas médicas, a informática médica se transformou gradualmente um campo interdisciplinar teoricamente baseado em tecnologia e Ciência da Computação que engloba medicina e Ciência da Informação e gestão para alcançar a gestão médica digitalizada em nível global.

Marrone (2020) apresentou como objetivo principal em sua pesquisa propor a vinculação de entidades para identificar palavras-chave. Tais palavras-chave são noções significativas que representam o texto e são doravante denominados “tópicos”. Os tópicos extraídos representam o conteúdo das publicações. O intuito da pesquisa é a aplicação da representação da informação para identificar temas emergentes de pesquisas.

Miranker, Giordano (2020) em seu trabalho buscaram explorar maneiras de analisar comunicados sociais e de mídia da Patrulha de Alfândega e Fronteira dos Estados Unidos para entender a morte de migrantes na fronteira Texas-México. E para essa análise eles utilizaram os métodos: linguística de *corpus*, Representação Espacial Qualitativa, Tríplices Semânticas e Processamento de Linguagem Natural com abordagem em mineração de textos. Os autores destacaram o uso dos métodos de PLN e Linguística de *corpus* para área de forense linguística (o estudo de evidências linguísticas em um tribunal de justiça e, em geral, textos e discursos jurídicos). A linguística forense se “preocupa em criar perfis e automatizar análises de *big data* de textos não estruturados, como e-mails, blogs, documentos de textos e mídias sociais, com o objetivo de descobrir “relações e padrões em evidências forenses digitais” (MIRANKER, GIORDANO, 2020, p. 2).

Shah *et al.* (2020) buscaram desenvolver e comparar o desempenho de cinco definições de coorte para identificar pacientes com Fibrilação Atrial do Registros Médicos Eletrônicos. Eles utilizaram a abordagem de mineração de textos. Os autores apontaram que a mineração de textos com o PLN aproveita a narrativa não estruturada dos cuidados de rotina para identificar coortes de pacientes, o que poderia auxiliar a seleção mais precisa de pacientes.

Yang *et al.* (2021) em sua pesquisa exploraram três modelos baseados em transformador amplamente utilizados (ou seja, BERT, RoBERTa e XLNet) para extração de relações clínicas usando dois conjuntos de dados de referência. Os autores citaram que a extração de relações clínicas é uma tarefa fundamental do PLN para identificar relações entre conceitos clínicos a partir de narrativas clínicas, o que pode construir perfis abrangentes de pacientes.

Cernile *et al*. (2020) tiveram como objetivo principal em sua pesquisa, demonstrar a viabilidade de usar uma abordagem de gráfico de rede para navegação rápida da literatura COVID-19 em um formato disponível publicamente para auxiliar na descoberta de conhecimento. Os autores utilizaram as abordagens de gerenciamento de conhecimento (mineração de textos e redes gráficas) que podem efetivamente permitir a navegação rápida e a exploração de inter-relações de entidades para melhorar a compreensão de doenças como a COVID-19.

Du, Chen, Zhang (2021) propuseram uma abordagem para mapear a interação entre diferentes entidades de conhecimento usando a estrutura em árvore de *Medical Subject Headings* para obter insights sobre as interações entre oferta de informática, demanda de saúde e aplicações tecnológicas em inovação médica digital na China. Du, Chen, Zhang (2021) afirmaram que a informática médica, converteu-se em uma disciplina científica estabelecida em todo o mundo. Estuda os dados, informações e conhecimentos da biomedicina e da saúde e sua organização sistemática, representação e métodos de análise.

Jiang *et al.* (2021) buscaram capturar e transformar dados da análise informatizada sobre o Sistema Nacional de Avaliação em Saúde para avaliar resultados do dispositivo de ablação cardíaca.

Lei, Xu, Zhou (2021) neste estudo tiveram como objetivo revelar as características compartilhadas e distintas de duas comunidades de saúde on-line populares na China, analisando de forma sistemática e abrangente seu conteúdo gerado pelo usuário e os comportamentos dos usuários associados. Eles adotaram a técnica de mineração de textos para melhor entender e prever a participação do usuário. Este artigo utilizou a técnica de mineração de textos para realizar a análise de redes sociais, com abordagem de análise de Tópicos (que tem objetivo de extrair tópicos conceituais, determinar seus tipos e analisar suas estruturas internas latentes em um grande *corpus* de texto).

Mesquita, Dias, Souza (2021) buscaram em sua pesquisa a extração

automática de Sintagma Nominal (SN), utilizando a tarefa de Extração de Informação. Os autores argumentaram que com o crescente volume informacional e juntamente com os avanços na área de PLN, abre espaço para novas pesquisas. “Uma delas estaria no uso de sintagmas nominais em sistemas de recuperação da informação, como a indexação automática”. (MESQUITA, DIAS, SOUZA, 2021, p. 2).

Souza, Felipe (2021) em sua pesquisa tiveram como objetivo a aplicação de técnicas de Mineração de texto e de Extração de Informações em anamneses de prontuário eletrônico do paciente para extração de termos visando enriquecimento de Terminologia clínica do tipo ontologia. Os autores citaram que a sua pesquisa é conduzida no âmbito da Ciência da Informação, é uma forma de iniciativa de aplicação de PLN com vistas a recuperação da informação do campo médico da ginecologia. “A CI, nas tarefas de organização do conhecimento, atua em domínios como da Medicina buscando soluções para os problemas de informação e para a melhor gestão dos recursos em saúde” (SOUZA, FELIPE, 2021, p. 53).

Andrushchenko *et al*. (2022) em seu artigo propuseram a criação de dois conjuntos de dados (corpora) que foram gramaticalmente investigados para o objetivo de analisar as características linguísticas para responder questões de humanidades. Para responder essas questões os autores utilizaram a técnica de PLN, focando inicialmente o processo de organização e limpeza dos dados e também no desenvolvimento de um sistema de busca para facilitar a utilização do PLN.

4.1.5 Categoria de Modelagem de tópicos e Classificação de textos

Nessa categoria são apresentados como os autores realizam aplicações em Modelagem de tópicos e Classificação de texto em suas pesquisas.

Quadro 8 - Artigos selecionados na categoria de Modelagem de tópicos e Classificação de texto

TÍTULOS	AUTORES	ANO
Informatics can identify systemic sclerosis (SSc) patients at risk for scleroderma renal crisis	REDD, Doug *et al.*	2014
Trends in biomedical informatics: automated topic analysis of JAMIA articles	HAN, Dong *et al.*	2015
Assistente de conhecimento conceitual como um sistema intencional para processos tutoriais em educação a distância	MEDEIROS, Luciano Frontino de; MOSER, Alvino; SANTOS, Neri dos	2015
Evaluating topic model interpretability from a primary care physician perspective	ARNOLD, Corey W. *et al.*	2016

Latent topics resonance in scientific literature and commentaries: evidences from natural language processing approach	WANG, Tai *et al.*	2018
Toward a clinical text encoder: pretraining for clinical natural language processing with applications to substance misuse	DLIGACH, Dmitriy; AFSHAR, Majid; MILLER, Timothy	2019
Avaliação das etapas de pré-processamento e de treinamento em algoritmos de classificação de textos no contexto da recuperação da informação	GUIMARÃES, Lucas M. S.; MEIRELES, Magali R. G.; ALMEIDA, Paulo Eduardo M.	2019
Coding and classifying GP data: the POLAR project	PEARCE, Christopher *et al*.	2019
Using the contextual language model BERT for multi-criteria classification of scientific articles	AMBALAVANAN, Ashwin Karthik; DEVARAKONDA, Murthy V.	2020
Padrões emergentes e tendências da estrutura científica internacional no domínio "discurso do ódio"	PUERTA-DíAZ, Mirelys; OVALLE-PERANDONES, María-Antonia; MARTÍNEZ-ÁVILA, Daniel	2020
Automatic Taxonomy Classification by Pretrained LanguageModel	KUWANA, A.; OBA, A.; SAWAI, R.; PAIK, I.	2021

Fonte: Autoria própria, 2023

Dos artigos analisados, 11 foram classificados nesta categoria, sendo 3 localizados a partir da Base de Dados BRAPCI e 8 a partir da Base de Dados SCOPUS. Abaixo, apresentamos uma análise sucinta de cada artigo selecionado:

Redd (2014) em seu trabalho utilizou um classificador de documentos baseado em Máquina de Vetor de Suporte (SVM) para identificar potenciais pacientes com esclerose sistêmica no Veterans Health Administration que estavam em uso de prednisona.

Han *et al.* (2015) buscaram descobrir quais as tendências em informática biomédica, para isso, utilizaram tensores (ou seja, matrizes multidimensionais) para representar a interação entre tópicos, tempo e citações e decomposição tensorial aplicada para automatizar a análise.

Medeiros, Moser, Santos (2015) em sua pesquisa descreveram o Assistente de Conhecimento Conceitual (ACC), uma ferramenta com arquitetura multi-agente, construída para interação com o usuário através de PLN. Medeiros, Moser, Santos (2015) explicaram que o ACC utiliza técnicas de busca aproximadas de termos a partir das perguntas fornecidas, o modelo de perguntas e respostas.

Arnold *et al*. (2015) buscaram avaliar a coerência dos tópicos e sua capacidade de representar o conteúdo de relatórios clínicos. Esses tópicos foram gerados pela técnica de Alocação Latente de Dirichlet, no qual refere se a um modelo de *bag-of-words* em que os documentos são misturas de tópicos geradores de

palavras. Os autores destacaram que, o LDA propõe um modelo generativo para misturas de tópicos de documentos usando um Dirichlet anterior à distribuição de tópicos de um documento.

Da mesma forma, Wang *et al*. (2018) utilizaram a técnica de Alocação Latente de Dirichlet em sua pesquisa, na qual eles apontaram que como uma das abordagens de modelagem de tópicos específicos, a LDA representa cada documento de uma coleção como uma mistura limitada sobre um conjunto de tópicos, cuja distribuição é assumida como tendo um esparso a priori de Dirichlet verificaram que os tópicos extraídos pela LDA são os mais próximo dos julgamentos humanos, comparando com a indexação semântica latente probabilística e o modelo de tópicos correlacionados.

Dligach, Afshar, Miller (2019) tiveram como objetivo principal em seu estudo desenvolver algoritmos para codificar texto clínico em representações que possam ser usadas para uma variedade de tarefas de fenotipagem. Dligach, Afshar, Miller (2019) apontaram que as representações de texto oriundas deste codificador, quando usadas para tarefas de aprendizado de máquina *downstream*, provavelmente beneficiarão o desempenho do classificador porque têm poder de representação de um grande conjunto de dados. Quanto melhor a representação, melhor a recuperação da informação.

Guimarães, Meireles, Almeida (2019) propuseram uma avaliação quantitativa das etapas de pré-processamento e de treinamento de um classificador no processo de classificação automática de dados não estruturados. Em seu experimento, eles utilizam a aplicação da técnica de análise de sentimento e avaliam o treinamento do classificador. Guimarães, Meireles, Almeida (2019), apontaram que esse estudo poderá contribuir com os trabalhos de pesquisa relacionados à aprendizagem de máquina voltada para os processos de organização e de recuperação da informação. Diante disso, podemos mencionar Ladeira (2010) que ressalta a importância do entendimento das técnicas de PLN para serem aplicadas na Ciência da Informação, nos contextos de representação e de recuperação da informação.

Pearce *et al.* (2019) investigaram em seu estudo um método de codificar e classificar automaticamente dados clínicos geral. O método desenvolvido consistiu na utilização do PLN para codificar automaticamente os textos, seguido de um processo manual de correção de códigos. Pearce *et al.* (2019) afirmaram que a codificação os tornará clinicamente útil para agregação para fins de pesquisa e saúde da população.

Ambalavanan, Devarakonda (2020) em sua pesquisa argumentaram a

dificuldade em encontrar artigos científicos específicos em uma grande coleção, é um importante desafio de Processamento de Linguagem Natural no domínio biomédico. Em seu estudo os autores enquadraram o problema como classificação de texto, e para isso, eles realizaram vários experimentos para comparar arquiteturas de ensemble.

Puerta-Díaz, Ovalle-Perandones, Martínez-Ávila (2020) tiveram como principal objetivo em sua pesquisa identificar padrões e tendências da estrutura científica internacional sobre discurso de ódio. Eles utilizaram no PLN a técnica de Alocação Latente de Dirichlet (LDA), “esse método de modelagem probabilística de tópicos foi aplicado para o conteúdo dos campos títulos, resumos e palavras-chave para a extração e cálculo da frequência de ocorrência das palavras abordadas nas publicações” (PUERTA-DÍAZ, OVALLE-PERANDONES, MARTÍNEZ-ÁVILA, 2020, p. 967).

Kuwana, Oba, Sawai, Paik (2021) propuseram em sua pesquisa a classificação automática de ontologias por modelo pré-treinado. Kuwana, Oba, Sawai, Paik (2021, p. 1) destacaram que “nos últimos anos, a geração automática de ontologias tem recebido atenção significativa na Ciência da Informação como um meio de sistematizar grandes quantidades de dados on-line”.

4.2 A Inter-Relação do Processamento de Linguagem Natural com a Ciência da Informação

Nesta seção vamos discutir quais as inter-relações entre o campo do Processamento de Linguagem Natural com a área da Ciência da Informação, e destacando as contribuições entre as temáticas.

Dentro da categoria de **fundamentação e conceituais** foi possível identificar o surgimento do termo de PLN nas pesquisas acadêmicas dentro da área da CI nas publicações registradas a partir de 1995, fruto da revolução técnico-cientifica, que nas argumentações de Saracevic (1996) afirma a inter-relação da CI com outras disciplinas.

Ainda dentro dessa categoria foi possível identificar quais as tarefas de PLN estão sendo pesquisadas e para quais finalidades dentro da área da CI, como: para o mapeamento cruzado automático de terminologia de enfermagem, a mineração de patentes de nanomateriais, a tarefa de ligação de entidade, o uso do campo de PLN

para o uso no campo dos estudos métricos da informação, para a comparação de sistemas de indexação automática por atribuição, e na pandemia do Covid-19, foi possível observar a enorme contribuição do campo de PLN ao extrair informações úteis para a comunidade científica ao melhorar a recuperação da informação ao identificar produções de alto impacto para a área médica.

Na categoria de **Pré-processamento de texto**, essa etapa é a preparação dos dados de texto ou fala. Várias aplicações de PLN passa por essa preparação. E é nessa preparação que destacamos a relação da Linguística com o Processamento de Linguagem Natural.

Como foi mencionado anteriormente nesse trabalho sobre as análises linguísticas, a análise morfológica é utilizada nas etapas de lematização e stemização, que trabalha com a estrutura da palavra. Para a recuperação da informação essa etapa se destaca na extração de radicais (*stemming*), para a substituição de uma palavra normalizada.

Assim como a morfologia, a sintaxe também estuda a estrutura das palavras, só que em sua composição em uma sentença. De forma bem simples, em formato de uma pequena árvore sintática. Por meio da análise sintática podemos identificar a estrutura sintagma nominal, que podem ser extraídas automaticamente com o PLN, e o seu uso em sistemas de recuperação da informação, como forma de indexação automática.

Nessa categoria ainda, foi constatado que a aplicação que mais destacou foi a sumarização automática de textos, que tem o objetivo de transmitir ou comunicar somente o que é importante de uma fonte textual de informação.

Para a área da Ciência da Informação, essa aplicação pode contribuir para uma melhoria da representação, mediação, recuperação e uso da informação, como por exemplo, nas páginas da *Web*, prover de um importante metadado: o resumo.

Como já visto anteriormente, o resumo está inserido no processo de Análise Documentária, dentro da categoria de representação temática da informação. A autora Cordeiro (2019) ressalta que no resumo documentário é realizado a condensação e a representação dos argumentos principais dos documentos, mas, ao contrário da indexação, nele, a descrição é produzida de forma narrativa, ou seja, é realizada a sumarização narrativa dos conteúdos (temáticos) dos documentos.

Podemos realizar algumas reflexões sobre quais as possíveis contribuições da área da CI para essa aplicação.

Segundo Souza *et al*. (2017), o processo de sumarização automática de textos realiza uma etapa semelhante ao processo de indexação automática, o grande desafio é, conforme Lancaster (2004), a extração de termos representativos do conteúdo dos documentos.

Ainda os autores supracitados, o processo de indexação por extração automática é baseado unicamente em critérios estatísticos, e por isso, apresenta limitações. Na indexação por atribuição automática, esse processo contém uma preocupação quanto aos aspectos semânticos do texto dos documentos. A solução é agregar outros conceitos aos termos (a partir da utilização de um instrumento de controle terminológico), ampliando a capacidade de representação temática do conteúdo do documento e somando novo valor à indexação automática feita em primeira instância. Então a combinação das técnicas de indexação automática por extração e por atribuição, é capaz de elaborar um texto de menor dimensionalidade.

Diante disso, podemos concluir que os estudos consolidados no campo de organização e representação da informação e do conhecimento também contribui com essa tarefa de PLN, com o uso de controle terminológico, enriquecimento semântico.

Dentro da categoria da **Análise semântica e Representação de textos**, novamente identificamos a relação com a Linguística. Na área da Linguística, a semântica busca estudar o significado das palavras, de acordo com o contexto inserido. Essa técnica é fundamental (e um pouco mais complexa) em PLN, por envolver o entendimento real da linguagem. Então, ensinar para a máquina quando identificar o significado da palavra em certa sentença demanda mais dados e estudos.

Na etapa da análise semântica, ocorre a resolução da ambiguidade de palavras, pois tais ambiguidades muitas vezes só podem ser solucionadas no contexto de uma frase ou um parágrafo.

Nessa categoria as tarefas que mais se destacou foram a aplicação de similaridade semântica, que é baseada no conceito de proximidade semântica, que estabelece medida entre dois termos que se refere ao quão similar são os seus respectivos sentidos (significados) em um dado contexto. Existem vários algoritmos que se propõe a calcular a similaridade semântica. A medida de similaridade semântica baseada em ontologia “abrange todas as abordagens que usam recursos e bases de conhecimento (como ontologias, dicionários e vocabulários) para melhorar o cálculo do grau de similaridade semântica entre os termos” (SILVA, 2008, p. 20).

Para CI podemos citar a aplicação de medida de similaridade para o uso de

modelos alternativos para a avaliação da ciência como: indicadores de produção e desempenho científico (OLIVEIRA, DIAS, PINTO, 2019), recuperação da informação para aplicação em estratégias no mercado de medicina diagnóstica (BRANQUINHO, PORTO, ALMEIDA, 2015) e para avaliar a similaridade entre os documentos (MEIRELES, CENDÓN, ALMEIDA, 2016). A similaridade semântica pode contribuir significativamente para os processos de categorização e de classificação da informação, que são utilizados como ferramentas de apoio à recuperação de informação.

E como mencionado anteriormente, a similaridade semântica utiliza de ontologias, dicionários e vocabulários controlados, instrumentos de controle terminológico que a área da CI tem grande autoridade sobre o assunto. Diante disso, podemos identificar a contribuição de uma das subáreas da CI: Organização e Representação do Conhecimento.

Outra aplicação em destaque é a de Relacionamento semântico, que segundo Ballatore, Bertolotto, Wilson (2014) às medidas computacionais de Relacionamento semântico desempenham um papel fundamental no Processamento de Linguagem Natural, recuperação de informações (RI) e desambiguação de sentido de palavras, fornecendo acesso a conexões semânticas mais profundas entre palavras e conjuntos de palavras. Podemos destacar que o Relacionamento semântico pode ser utilizado para a área da CI, na desambiguação da relação em tesauro e o seu reuso em ontologias (MACULAN, AGANETTE, 2017), na extração de relacionamentos semânticos para o enriquecimento de ontologia de domínio (SHEN *et al*. 2012). As Relações semânticas são fundamentais para a compreensão de representações da realidade que envolvem estruturas conceituais.

Dentro do âmbito da CI, a primeira referência sobre relações semânticas entre conceitos, encontrada na literatura pesquisada, remete-se à Dahlberg (1978), em sua Teoria do Conceito, amplamente utilizada na Biblioteconomia e Ciência da Informação, ela classifica os relacionamentos em lógicos e semânticos. As relações lógicas são baseadas nas características comuns entre os conceitos. De acordo com a autora, esse tipo de relação é muito importante, pois, a partir dele, é possível comparar os conceitos, organizá-los e relacioná-los semanticamente. Já as relações semânticas sugeridas por Dahlberg são: hierárquica, partitiva, de oposição e funcional.

Na quarta categoria apresentamos as aplicações de **Extração de**

Informações (EI) e Mineração de textos, ambas aplicações passam pelo o processo de pré-processamento de texto. As técnicas de EI podem ser parte da tarefa de mineração de textos para facilitar a extração do conhecimento.

A aplicação de Extração de Informações consiste em associar logicamente as entidades previamente identificadas, reconhecidas e classificadas" (ARANHA; PASSOS, 2006, p. 6).

Para a área da CI, os sistemas de extração de informações também permitem melhorar o desempenho de sistemas de recuperação de informações através da integração e sintetização da informação, evitando desta forma a ocorrência de redundâncias em textos que tratam do mesmo assunto (BATRES *et al*., 2005).

Segundo Wives e Loh (1999) a EI foi uma evolução natural da área de recuperação de informações. Grishman (1997) afirma que a extração de informações envolve a criação de uma representação estruturada da informação selecionada e extraída do texto. Podemos concluir que a tarefa de Extração de informações contribui significativamente para área da recuperação da informação.

Para a aplicação de Extração de informações o uso de ontologias pode permitir uma ampliação da extração de informações ao fornecer um sistema conceitual expresso por um conjunto de termos e suas relações que permitem, a partir de um determinado termo, a localização de termos mais amplos ou mais genéricos, sinônimos, termos oposto e termos associados em geral.

A aplicação de Mineração de textos consiste em extrair regularidades, padrões ou tendências de grandes volumes de textos em linguagem natural, normalmente para objetivos específicos, tal como a tomada de decisões (ARANHA; PASSOS, 2006).

Segundo Limiro, Silva, Cordeiro (2022), a mineração de textos pode ser compreendida como uma subárea da Recuperação da Informação (RI), na qual, através de um conjunto de rotinas de processamento e análise de padrões, a informação é recuperada a partir de dados textuais, gerando, consequentemente, o conhecimento. Fan *et al.* (2006) descrevem algumas tarefas de mineração de textos: perguntas e respostas, sumarização, extração de informação, categorização, agrupamento, visualização de informação e rastreamento de informação, muitas dessas aplicações são realizadas dentro da área da CI.

Alguns exemplos de aplicação de mineração de texto no âmbito da CI são: Mineração de textos aplicada a postagens do Twitter sobre Coronavírus (AFONSO,

GOTTSCHALG-DUQUE, 2020), Direitos autorais e mineração de dados e textos no combate à Covid-19 no Brasil (SOUZA, SCHIRRU, ALVARENGA, 2020), Mineração de textos para agrupamento de teses e dissertações por meio de análise de similaridade (LIMIRO, SILVA, CORDEIRO, 2022), Mineração de texto e clusterização em estudos bibliométricos (PEREIRA *et al.*, 2022) e Recuperação e classificação de sentimentos de usuários do Twitter em período eleitoral (MATOS, MAGALHÃES, SOUZA, 2020). Podemos identificar que a aplicação de mineração de textos para a área da CI, contribui para busca de informações em documentos, identificação de termos mais relevantes de um documento, resumir documentos, realizar análises qualitativas ou quantitativas em documentos de texto e assim, auxiliar na recuperação da informação. Para essa aplicação, o uso de ontologias contribui significativamente para a redução de dados a serem minerados.

Na última categoria, apresentamos a **Modelagem de tópicos e Classificação de textos**. Na tarefa de modelagem de tópicos é utilizado o método de Alocação Latente de Dirichlet (LDA). Essa aplicação é um dos métodos mais populares para executar a modelagem de tópicos. É usada no Processamento de Linguagem Natural para encontrar textos semelhantes.

Para a área da Ciência da Informação, essa aplicação contribui para classificação de documentos em grupos e reconhecimento do assunto principal e secundário em cada grupo de documentos. Também auxilia para a organização e sumarização de coleções, sistemas de busca e recomendação e detecção de conteúdo duplicado. Essa aplicação contribui com a organização e representação da informação.

Segundo Souza, Souza (2019, p. 3) “o modelo Latent Dirichlet Allocation (LDA) é um dos modelos generativos mais utilizados para organizar grandes coleções de documentos”. Trata-se de um modelo que utiliza uma abordagem bayesiana, no qual os documentos contidos em um corpus são representados como uma mistura aleatória de tópicos latentes que emergem por meio de uma estrutura não supervisionada. Cada tópico é caracterizado por uma distribuição de palavras e pesos que compreendem cada um dos documentos.

Podemos resumir que o método LDA realiza a escolha do tópico e da palavra aleatoriamente. Mas mostra-se limitado, quando se deseja descobrir as relações entre palavras e documentos com o mínimo de intervenção humana. Nessa situação a CI coopera com a abordagem de ancoragem de palavras. Segundo Fernandes (2016)

essas palavras são introduzidas como âncoras de conhecimento de domínio de maneira semelhante a técnicas como o método do gargalo da informação.

A Classificação automática de documentos de texto utiliza diferentes técnicas de aprendizado de máquina aplicadas a coleções diferentes de documentos, procurando extrair padrões relevantes para organizar as informações do texto em um formato que facilite o processo de Recuperação da Informação (ARANHA; PASSOS, 2006). A classificação de textos é a tarefa de associar textos em linguagem natural a rótulos pré-definidos.

Para área da CI, a classificação automática de textos pode ser aplicada em uma grande variedade de contextos como, por exemplo: indexação automática de textos, identificação de autores de textos, classificação hierárquica de páginas da Internet e geração automática de metadados. Podemos perceber que essa tarefa contribui com a representação e recuperação da informação.

Para essa aplicação, a contribuição da CI, é na redução da dimensionalidade em textos pela seleção de termos, é uma das técnicas mais utilizadas para esse fim, é o uso de tesauro. Que tem o objetivo de reduzir os termos com significados similares (sinônimos), relacionados e opostos (antônimos).

5 CONSIDERAÇÕES FINAIS

Considerando o aumento das informações que são geradas em documentos no formato textuais, especialmente, em ambientes digitais, podemos observar a dificuldade dos usuários em suprir as suas necessidades informacionais, pois com o aumento e acesso às informações rapidamente, trouxeram problemas no acesso e recuperação de documentos nos sistemas de informações documentais, e com isso fazendo que o usuário se sinta sobrecarregado e perdido diante desse volume de dados e informações.

Diante disso, torna-se inviável a manipulação manual das informações, tornando-se necessário a utilização de mecanismos e ferramentas que possibilitem o processamento automático de informações. A principal abordagem de análise de texto e linguagem por meio computacional é chamada de Processamento de Linguagem Natural.

A presente pesquisa fez um levantamento de artigos referente a Ciência da Informação e o Processamento de Linguagem Natural existentes em duas bases de dados (BRAPCI e SCOPUS) em um intervalo de dez anos, ou seja, de 2012 a 2022. Foram recuperados 81 artigos, sendo que 68 foram analisados devido à repetição de alguns artigos nas duas bases ou não terem informações. Com a análise dos artigos, foi possível identificar alguns fatores:

Foi possível identificar quais tarefas de PLN foram mais utilizadas nos artigos analisados. A tarefa mais aplicada é a Extração de Informações. Em vários artigos mais de uma tarefa de PLN foi utilizada.

Dentre as tarefas de PLN, é possível identificar aquelas com forte correlação com a Ciência da Informação, como sumarização automática de textos, medida de similaridade semântica, medida de relacionamento semântica, classificação automática de textos, mineração de textos, extração de informações, *Question Answering* e *clustering*.

Outro ponto que chamou a atenção, foi a identificação da área de conhecimento que mais utilizaram o campo de PLN. Como principal usuária de PLN, pode-se citar a área médica.

Entre as aplicações na área médica estão tarefas como registros médicos eletrônicos, para atividade clínica, detecção de atributos de conceitos médicos no texto clínico, extração de informação sobre a pandemia de Covid-19, extração de

informação sobre mortalidade por câncer, extração de relação de tratamento, extração de informação em mamografia, identificação de potenciais pacientes com esclerose sistêmica, higienização de dados sensíveis médicos, codificação de texto clínico para tarefas de fenotipagem, reconhecimento automático de achados negados em laudos radiológicos, sumarização de textos para dados clínicos, e outros. Esse resultado aponta que a área médica tem buscado utilizar mais os recursos de PLN para otimizar suas atividades operacionais e de pesquisa.

Após as análises dos artigos foi possível observar as contribuições do campo de Processamento de Linguagem Natural para a área da Ciência da informação, essas contribuições são mais significativas para a área de recuperação da informação, pois várias aplicações têm a sua finalidade de extrair informações úteis para o seu usuário.

Além disso, foi possível identificar que a área da Ciência da Informação contribui para o campo de Processamento de Linguagem Natural, com o uso de ontologias, taxonomia, tesauros, dicionários, indexação e organização e representação da informação e do conhecimento.

Da mesma forma, foi identificado as contribuições da Linguística e Terminologia para o campo de Processamento de Linguagem Natural, com o uso das análises linguísticas e dos termos.

Diante disso, foi possível analisar a inter-relações do campo de Processamento de Linguagem Natural com a área da Ciência da Informação, na qual as duas temáticas conseguem usufruir das técnicas e especificidades da outra.

Perante o exposto, a Ciência da informação pode e deve usufruir das ferramentas computacionais desenvolvidas no âmbito das pesquisas em PLN, empregando-as nos processos de catalogação e seguido da recuperação nos centros informacionais, assim como na realização de modelos de representação de informação.

Finalmente, vale destacar que no escopo desta dissertação não se esgotaram as possibilidades de discussão e ainda há muito que ser explorado. Pois o campo de PLN está em desenvolvimento constante. Essas inovações têm grande possibilidades de contribuir de forma construtiva com a Ciência da Informação, proporcionando a produção de novos conhecimentos.

REFERÊNCIAS

AFONSO, A. R.; GOTTSCHALG-DUQUE, C. Mineração de textos aplicada a postagens do twitter sobre coronavírus: uma análise na linha do tempo. **Liinc em revista**, v. 16, 2020. DOI: 10.18617/liinc.v16i2.5325 Acesso em: 04 maio 2023

ALMEIDA, B. Terminologia e organização do conhecimento: linguagens, vocabulários e sistemas. **Linha D'Água**, [S. l.], v. 34, n. 2, p. 26-46, 2021. DOI: 10.11606/issn.2236-4242.v34i2p26-46. Disponível em: https://www.revistas.usp.br/linhadagua/article/view/188258. Acesso em: 29 ago. 2022.

ALMEIDA, G. M. B. O percurso da Terminologia: de atividade prática à Consolidação de uma disciplina autônoma. **Tradterm**, 9, 2003. p. 211-222. Disponível em: https://doi.org/10.11606/issn.2317-9511.tradterm.2003.49087 Acesso em 11 ago. 2022.

ALVARES, L. M. A. R. **Linguagens Documentárias**. Faculdade de Ciência da Informação, Universidade de Brasília. jun. 2011. Apresentação em PowerPoint. 52 slides. color. Disponível em: http://lillianalvares.fci.unb.br/phocadownload/Analise/Metodos/Aula41LD.pdf. Acesso em: 17 ago. 2022.

AMARAL, B. M.; SILVA, E. M. S.; ALMEIDA, A. M. G. Análise de Sentimentos/Mineração de Opinião: uma revisão bibliográfica. **RETEC**, Ourinhos, v. 10, n. 2, p. 80-99, jan./jun. 2017. Disponível em: https://www.fatecourinhos.edu.br/retec/index.php/retec/article/view/256. Acesso em: 25 jul. 2022.

AMARAL, D. O. F. **O reconhecimento de entidades nomeadas por meio de conditional Random Fields para a língua portuguesa**. 2013. Dissertação (Mestrado em Ciência da Computação) - Pontifícia Universidade Católica do Rio Grande do Sul, Porto Alegre, 2013. Disponível em: https://hdl.handle.net/10923/5772. Acesso em: 25 jul. 2022.

ARANHA, C.; PASSOS, E. A Tecnologia de Mineração de Textos. **Revista Eletrônica de Sistemas de Informação**, Curitiba, v. 5,n. 2, p.1-8. 2006. Disponível em: http://www.periodicosibepes.org.br/index.php/reinfo/article/view/171/66 Acesso em: 04 maio 2023.

ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS. **NBR 12676**: Métodos para análise de documentos – Determinação de seus assuntos e seleção de termos de indexação. Rio de Janeiro: ABNT, 1992. Disponível em: https://www.semas.pa.gov.br/wp-content/uploads/2018/04/NBR-12676-INDEXACAO.pdf. Acesso em: 07 set. 2022.

BAEZA-YATES, R.; RIBEIRO-NETO, B., **Modern Information Retrieval**. Addison-Wesley, 1999. Disponível em: http://web.cs.ucla.edu/~miodrag/cs259-security/baeza-yates99modern.pdf. Acesso em: 26 jul. 2022.

BALLATORE, A.; BERTOLOTTO, M.; WILSON, D. C. An Evaluative Baseline for Geo-Semantic Relatedness and Similarity. **GeoInformatica**, v. 18, n. 4, 2014, p. 747-767. Disponível em: https://www.scopus.com/inward/record.uri?eid=2-s2.0-84908118044&doi=10.1007%2fs10707-013-0197-8&partnerID=40&md5= DOI: 10.1007/s10707-013-0197-8. Acesso em: 22 jul. 2022

BARANOW, U. G. Perspectivas na contribuição da linguística e de áreas afins à ciência da informação. **Ciência da Informação**, [S. l.], v. 12, n. 1, 1983. Disponível em: https://revista.ibict.br/ciinf/article/view/191. Acesso em: 22 set. 2022.

BARDIN, L. **Análise de conteúdo**. Rio de Janeiro: Edições 70, 2011.

BARRETO, A. de A. Uma história da ciência da informação. In. TOUTAIN, L. M. B. B. (Org.). **Para entender a Ciência da Informação.** Salvador: EDUFBA, 2007, p. 13-34. Disponível em: <https://repositorio.ufba.br/ri/bitstream/ufba/145/1/Para%20entender%20a%20ciencia%2 0da%20informacao.pdf>. Acesso em: 20 maio 2023.

BARROS, F. A.; ROBIN, J. **Processamento de Linguagem Natural** (JAI/SBC'96). 1996. (Desenvolvimento de material didático ou instrucional - Textos Didáticos). Disponível em: https://www.cin.ufpe.br/~fab/cursos/jai96/ProcessamentoDeLinguagemNatural.pdf. Acesso em: 02 ago. 2022.

BATRES, E. J. Q.; OLIVEIRA, A. P.; GABRIELLI, B. V.; AMORIM, V. P.; MOREIRA, A. Uso de ontologias para a extração de informações em atos jurídicos em uma instituição pública. **Encontros Bibli: Revista Eletrônica de Biblioteconomia e Ciência da Informação,** v. 10, n. 19, p. 73-88, 2005. DOI: 10.5007/1518-2924.2005v10n19p73 Acesso em: 04 maio 2023

BOCCATO, V. R. C. A linguagem documentária vista pelo conteúdo, forma e uso na perspectiva de catalogadores e usuários. *In*: FUJITA, M. S. L. (org.) *et al*. **A indexação de livros**: a percepção de catalogadores e usuários de bibliotecas universitárias. Um estudo de observação do contexto sociocognitivo com protocolos verbais [online]. São Paulo: Editora UNESP; São Paulo: Cultura Acadêmica, 2009. 149 p. Acesso em: https://books.scielo.org/id/wcvbc/pdf/boccato-9788579830150-08.pdf. Acesso em: 22 ago. 2022.

BORKO, H. Information Science: What is it? **American Documentation**, v.19, n.1, p.3-5, Jan. 1968. (Tradução Livre). Disponível em: https://edisciplinas.usp.br/pluginfile.php/1992827/mod_resource/content/1/Borko.pdf. Acesso em: 21 set. 2022.

BRANQUINHO, L. P.; PORTO, R. M. A. B.; ALMEIDA, M. B. Modelo para suporte à descoberta de conhecimento em base de dados (kdd): aplicação em estratégias no mercado de medicina diagnóstica. **Pesquisa Brasileira em Ciência da Informação e Biblioteconomia**, v. 10, n. 2, 2015. Disponível em: http://hdl.handle.net/20.500.11959/brapci/30488. Acesso em: 02 maio 2023.

BROOKSHEAR, J. G. **Ciência da Computação**: uma visão abrangente. 11. ed. Porto Alegre: Grupo A, 2013. E-book. ISBN 9788582600313. Disponível em: https://integrada.minhabiblioteca.com.br/#/books/9788582600313/. Acesso em: 21 set. 2022.

CABRÉ, M. T. La terminología hoy: concepciones, tendencias y aplicaciones. **Ciência da Informação**, [S. l.], v. 24, n. 3, 1995. DOI: 10.18225/ci.inf.v24i3.567. Disponível em: https://revista.ibict.br/ciinf/article/view/567. Acesso em: 16 ago. 2022.

CABRÉ, M. T. Una nueva teoría de la terminologia: de la denominación a la comunicación. In: ___. **La terminología, representacíon y comunicación**. Barcelona: Universitat Pompeu Fabra, IULA, 1999. p.109-127.

CARDOSO, O. N. P. Recuperação de Informação. **INFOCOMP Revista de Ciência da Computação**, [S. l.], v. 2, n. 1, p. 33–38, 2004. Disponível em: https://infocomp.dcc.ufla.br/index.php/infocomp/article/view/46. Acesso em: 26 jul. 2022.

CARRILHO JUNIOR, J. R. **Desenvolvimento de uma metodologia para mineração de textos**. 2007. Dissertação (Mestrado em Engenharia Elétrica) - Pontifícia Universidade Católica, Rio de Janeiro, 2007. Disponível em: https://www.maxwell.vrac.puc-rio.br/colecao.php?strSecao=resultado&nrSeq=11675@1 1>. Acesso em 22 ago. 2022.

CERVANTES, B. M. N. **A construção de tesauros com a integração de procedimentos terminográficos**. 2009. 209 f. Dissertação (Mestrado em Ciência da Informação) - Universidade Estadual Paulista, Marília, 2009. Disponível em: Acesso em: 28 ago. 2022. Disponível em: https://repositorio.unesp.br/handle/11449/103382. Acesso em: 02 set. 2022.

CERVANTES, B. M. N.; FUJITA, M. S. L.; RUBI, M. P. Terminologias em política de indexação. **Ibersid: revista de sistemas de información y documentación**, v. 2, n. issne 2174-081x; issn 1888-0967, p. 211-221, 2008. Disponível em: http://hdl.handle.net/20.500.11959/brapci/167017. Acesso em: 11 ago. 2022.

CINTRA, A. M. M.; TALAMO, M. F. G. M.; LARA, M. L. G.; KOBASHI, N. **Para entender as linguagens documentárias**. São Paulo: POLIS/APB, 1994. 72p.

COMARELLA, R. L.; CAFÉ, L. Chatterbot: conceito, características, tipologia e construção. **Informação & Sociedade: Estudos**, v. 18, n. 2, 2008. Disponível em: http://hdl.handle.net/20.500.11959/brapci/92052. Acesso em: 15 set. 2022.

CONEGLIAN, C. S. **Recuperação da informação com abordagem semântica utilizando linguagem natural:** a inteligência artificial na ciência da informação. Tese (Doutorado em Ciência da Informação) – Universidade Estadual de Paulista, Marília, 2020. Disponível em: http://hdl.handle.net/11449/193051. Acesso em: 07 mai. 2023.

CONEGLIAN, C. S.; SANTARÉM SEGUNDO, J. E. Inteligência artificial e ferramentas da web semântica aplicadas a recuperação da informação: um modelo conceitual com foco na linguagem natural. **Informação & Informação**, v. 27, n. 1, p. 625-651, 2022. Disponível em: https://ojs.uel.br/revistas/uel/index.php/informacao/article/view/44729. Acesso em: 24 ago. 2022.

CORDEIRO, R. I. N. **Processos e produtos de representação temática da informação**. Brasília, DF: CAPES: UAB; Rio de Janeiro, RJ: Departamento de Biblioteconomia, FACC/UFRJ, 2019. Disponível em: http://www.repositorio.bibead.ufrj.br/repbibead-verpdf.php?num=47&arquivo=Processos-e-Produtos-de-Representacao-Tematica-da-Informacao-LIVRO.pdf. Acesso em: 12 mai. 2023.

CHIOVATTO, M. Watson, uso de Inteligência Artificial (AI) e processos educativos em museus. **Revista Docência e Cibercultura**, [S.l.], v. 3, n. 2, p. 217-230, set. 2019. Disponível em: <https://www.e-publicacoes.uerj.br/index.php/re-doc/article/view/40293>. Acesso em: 04 out. 2022.

CHOWDHURY, G. C. Natural Language Processing. **Annual Review of Information Science and Technology**, v. 37, p. 51-89, 2003. Disponível em: https://doi.org/10.1002/aris.1440370103. Acesso em: 04 ago. 2022.

CRIVELLI, R. B. **Recuperação da informação por meio do processamento de linguagem natural.** 2011. Monografia (Bacharel em Sistemas de Informação) – Universidade Estadual do Norte do Paraná, Bandeirantes, 2011.

CUNHA, I. M. R. F; KOBASHI, N. Y. Análise documentária e inteligência artificial. **Revista Brasileira de Biblioteconomia e Documentação**, São Paulo, v. 24, n. 1/4, p. 38-62, 1991. Disponível em: https://www.brapci.inf.br/index.php/article/download/19184. Acesso em: 28 abr. 2023

DAHLBERG, I. Teoria do conceito. **Ciência da Informação**, [S. l.], v. 7, n. 2, 1978. DOI: 10.18225/ci.inf.v7i2.115. Disponível em: https://revista.ibict.br/ciinf/article/view/115. Acesso em: 20 mar. 2023

DENNING, P. J.; COMER, D. E.; GRIES, D.; MULDER, M. C.; TUCKER, A. B; TURNER, A. J.; YOUNG, P. R. Computing as a discipline. **Communication of the ACM**, v. 32, n. 1, p.9-23, 1989. Disponível em: https://dl.acm.org/doi/pdf/10.1145/63238.63239. Acesso em: 21 set. 2022.

DIAS, A. S. **Processamento de linguagem natural**. São Paulo: Platos Soluções Educacionais S.A, 2021. E-book. ISBN 9786589881995. Disponível em: https://integrada.minhabiblioteca.com.br/#/books/9786589881995/. Acesso em: 14 ago. 2022.

DIAS, C. A. Terminologia: conceitos e aplicações. **Ciência da Informação**, Brasília, v. 29, n. 1, p. 90-92, jan./abr. 2000. Disponível em: https://www.scielo.br/j/ci/a/yJhxDcM3VxH9DnwCfvzsCJP/?format=pdf&lang=pt. Acesso em: 02 ago. 2022.

DIKI-KIDIRI, M. Une approche culturelle de la terminologie. **Terminologies nouvelles**, n. 21, p. 27-31, 2000 (n. esp. Terminologie et diversité culturelle).

DUBOIS, D. Lexique(s) et catégories: de la percepction individuelle aux connaissances partagées. In:**Simpósio Internacional de Verano de Terminología**: Terminologia y cognición. Barcelona: IULA-UPF, 2001. p. 15-38.

FACHIN, O. **Fundamentos de metodologia**. 6. ed. São Paulo: Editora Saraiva, 2017. E-book. ISBN 9788502636552. Disponível em: https://integrada.minhabiblioteca.com.br/#/books/9788502636552/. Acesso em: 02 ago. 2022.

FALCÃO, L. C. J.; LOPES, B.; SOUZA, R. R. Absorção das tarefas de Processamento de Linguagem Natural (NLP) pela Ciência da Informação (CI): uma revisão da literatura para tangibilização do uso de NLP pela CI. **Em Questão**, Porto Alegre, v. 28, n. 1, p. 13–34, 2021. DOI: 10.19132/1808-5245281.13-34. Disponível em: https://seer.ufrgs.br/index.php/EmQuestao/article/view/111323. Acesso em: 12 jul. 2022.

FAN, W., WALLACE, L., RICH, S., ZHANG, Z. Tapping the power of text mining. **Communications of the ACM**, v. 49, n. 9, p. 76-82, 2006. Disponível em: <http://doi.acm.org/10.1145/1151030.1151032>. Acesso em: 15 maio 2023.

FERNANDES, N. E. N. Topic modeling: Latent Dirichlet Allocation vs Correlation Explanation alternative. **The Intelligence of Information**. 6 dez. 2016. Disponível em: https://theintelligenceofinformation.wordpress.com/2016/12/06/topic-modeling-latent-dirichlet-allocation-vs-correlation-explanation-alternative/. Acesso em: 01 mai. 2023.

FERNEDA, E. **Recuperação de Informação:** análise sobre a contribuição da Ciência da Computação para a Ciência de Informação. 2003. Tese (Doutorado em Ciência da Informação) – Universidade de São Paulo, São Paulo, 2003. Disponível em: https://www.teses.usp.br/teses/disponiveis/27/27143/tde-15032004-130230/publico/Tese.pdf. Acesso em: 24 ago. 2022.

FIORIN, J. L. (Org.). **Introdução à Linguística**: vol. 1 e 2. São Paulo: Contexto, 2003. Disponível em: https://docs.google.com/file/d/0BxE3XwG-HQyuSXNtMC1zNklHV2M/edit?resourcekey=0-74cYBrJyAvyaCVClXgQSvw. Acesso em: 24 ago. 2022.

FRANCELIN, M. M.; PINHO, F. A. Conceitos na organização do conhecimento. In: FRANCELIN, M. M.; PINHO, F. A. **Conceitos na organização do conhecimento**. Recife: Ed. Universitária da UFPE, 2011. p. 55-65.

FUSCHILO, C.; ALENCAR, A. J.; SCHMITZ, É. A. Proposta de uma Metodologia de Avaliação de Assistentes Virtuais Inteligentes: Aplicando Análise Envoltória de Dados (DEA) – Piperis. In: DESENHO DE PESQUISA - SIMPÓSIO BRASILEIRO DE SISTEMAS COLABORATIVOS (SBSC), 15., 2019, Rio de Janeiro. **Anais** [...]. Porto Alegre: Sociedade Brasileira de Computação, 2019. p. 34-39. Disponível em:

https://sol.sbc.org.br/index.php/sbsc_estendido/article/view/8348/8245. Acesso em: 03 set. 2022.

GABRIEL, M. **Inteligência Artificial**: do zero ao Metaverso. Rio de Janeiro: Grupo GEN, 2022. E-book. Disponível em: https://integrada.minhabiblioteca.com.br/#/books/9786559773336/. Acesso em: 23 jul. 2022.

GABRIEL, M. **Sem e Seo**: dominando o marketing de busca. 2. ed. São Paulo: Novatec Editora, 2012.

GARCIA, G. C. **Reconhecimento de Entidades Nomeadas na base de notificações de eventos adversos e queixas técnicas de dispositivos médicos no Brasil**. 2021. 158 f. Dissertação (Mestrado Profissional em Computação Aplicada) - Universidade de Brasília, Brasília, 2021. Disponível em: https://repositorio.unb.br/handle/10482/42718. Acesso em: 25 ago. 2022.

GAUDIN, F. **Pour une socioterminologie**. Rouen: Pub. Université de Rouen, 1993.

GERHARDT, T. E.; SILVEIRA, D. T. **Métodos de pesquisa**. Porto Alegre, RS: UFRGS, c2009. (Educação a distância (UFRGS Ed.) Disponível em: http://hdl.handle.net/10183/52806. Acesso em: 02 ago. 2022.

GIL, A. C. **Métodos e Técnicas de Pesquisa Social**. 7. ed. São Paulo: Atlas, 2019. E-book. ISBN 9788597020991. Disponível em: https://integrada.minhabiblioteca.com.br/#/books/9788597020991/. Acesso em: 02 ago. 2022.

GOMES, H. E.; CAMPOS, M. L. A. **A Organização do conhecimento na Web**: contribuições de Shiyali Ramamrita Ranganathan e de Ingetraut Dahlberg. Niterói: IACS-UFF, 2019. (Cadernos acadêmicos, 1). Disponível em: https://eooci.uff.br/wp-content/uploads/sites/319/2020/09/eooci_ca-1.pdf. Acesso em: 23 set. 2022.

GRISHMAN, R. Information extraction; techniques and challenges. In: INTERNATIONAL SUMMER SCHOOL SCIE-97, 1997, New York. **Proceedings…** New York: Springer-Verlag, 1997. p. 10-27.

GUINCHAT, C. MENOU, M. **Introdução geral às ciências e técnicas da informação e da documentação**. Brasília: MCT: CNPq: Ibict, 1994. Disponível em: http:/livroaberto.ibict.br/handle/1/1007. Acesso em: 16 ago. 2022.

ISO 1087-1. **Terminology work and terminology science** — Vocabulary. Génève: ISO, 2000. Disponível em: https://www.iso.org/standard/62330.html. Acesso em: 22 ago. 2022.

ISO 2788. **Documentation** — Guidelines for the establishment and development of monolingual thesauri. London: BSI, 1986. Disponível em: https://www.iso.org/standard/7776.html. Acesso em: 16 ago. 2022.

ISO 704. **Principles and methods of terminology.** 4.ed. Génève: ISO, 2022. Disponível em: https://www.iso.org/standard/79077.html. Acesso em: 22 ago. 2022.

JURAFSKY, D.; MARTIN, J. H. **Speech and Language Processing**. 3 ed. Pearson, 2020. Disponível em: https://web.stanford.edu/~jurafsky/slp3/ed3book.pdf. Acesso em: 23 ago. 2020

JIN, X.; WAH, B. W., CHENG, X.; WANG, Y. Significance and Challenges of Big Data Research**. Big Data Research**, Amsterdam, v. 2, n. 2, p. 59-64, 2015. https://doi.org/10.1016/j.bdr.2015.01.006. Acesso em: 24 ago. 2022.

KOBASHI, N. Y. Linguística textual e elaboração de informações documentárias: algumas reflexões. In: GASPAR, N. R.; ROMÃO, L. M. S. (Org.). **Discurso e texto**: multiplicidade de sentidos na Ciência da Informação. 1. ed. São Carlos: EduFscar, 2008, v. 1, p. 47-66.

KOBASHI, N. Y. Análise documentária e representação da informação. **Informare: Cadernos do Programa de Pós-Graduação em Ciência da Informação**, v. 2, n. 2, 1996. Disponível em: http://hdl.handle.net/20.500.11959/brapci/40976. Acesso em: 22 ago. 2022

LADEIRA, A. P. **Processamento de linguagem natural:** caracterização da produção científica dos pesquisadores brasileiros. 2010. Tese (Doutorado em Ciência da Informação) - Universidade Federal de Minas Gerais, Belo Horizonte, 2010. Disponível em: http://hdl.handle.net/1843/ECID-8B3Q6C. Acesso em: 24 ago. 2022.

LANCASTER, F. W. **Indexação e sumários:** teoria e prática. 2. ed. Brasília: Briquet de Lemos, 2004.

LARA, M. L. G**. Linguística Documentária**: seleção de conceitos. 2009. Tese (Livre Docência em Análise Documentária) – Universidade de São Paulo, São Paulo, 2009. doi:10.11606/T.27.2019.tde-21112019-191517. Disponível em: https://www.teses.usp.br/teses/disponiveis/livredocencia/27/tde-21112019-191517/publico//MarildaLopesGinesdeLaraLivreDocencia.pdf. Acesso em: 10 ago. 2022.

LARA, M. L. G. **Representação e Linguagens Documentárias:** Bases Teórico-Metodológicas. 1999. Tese (Doutorado em Ciência da Informação e Documentação) - Universidade de São Paulo, São Paulo, 1999. Disponível em: https://teses.usp.br/teses/disponiveis/27/27143/tde-02122019-153131/publico/MarildaLopesGinesdeLaraDoutorado.pdf. Acesso em: 05 ago. 2022

LARA, M. L. G. Novas relações entre Terminologia e Ciência da Informação na perspectiva de um conceito contemporâneo da informação. **DataGramaZero**, v. 7, n. 4, 2006. Disponível em: http://hdl.handle.net/20.500.11959/brapci/5938. Acesso em: 29 ago. 2022.

LARA, M. L. G. **Elementos de terminologia**. (apostila para uso didático). 2005. Disponível em: https://bibliotextos.files.wordpress.com/2012/03/elementos-de-terminologia.pdf. Acesso em: 23 ago. 2022.

LARA, M. L. G.; LIMA, V. M. A. **Introdução à Terminologia**. mai. 2015. Apresentação de PowerPoint. 42 slides. color. Disponível em: https://edisciplinas.usp.br/mod/resource/view.php?id=193541. Acesso em: 30 ago. 2022.

LIMA, G. N. B. Interfaces entre a ciência da informação e a ciência cognitiva. **Ciência da Informação**, v. 32, n. 1, 2003. Disponível em: https://revista.ibict.br/ciinf/article/view/1021. Acesso em:24 ago. 2022.

LIMA, V. M. A. **Terminologia, comunicação e representação documentária**. Dissertação (Mestrado em Ciência da Informação e Documentária) – Universidade de São Paulo, São Paulo, 1998. doi:10.11606/D.27.1998.tde-11052004-122839. Disponível em: https://www.teses.usp.br/teses/disponiveis/27/27143/tde-11052004-122839/publico/Term_Comum_Repres_Documentaria.pdf. Acesso em: 15 ago. 2022.

LIMA, V. M. A.; SANTOS, C. A. C. M.; MAIMONE, G. D. **A linguagem na perspectiva das Ciências da Linguagem**. 2017. Apresentação do Power Point. Disponível em: https://edisciplinas.usp.br/mod/resource/view.php?id=2036389&forceview=1. Acesso em 29 ago. 2022.

LIMIRO, R. M.; SILVA, N. R.; CORDEIRO, D. F. Mineração de textos para agrupamento de teses e dissertações por meio de análise de similaridade. **Revista Brasileira de Biblioteconomia e Documentação**, [S. l.], v. 18, p. 1–20, 2022. Disponível em: https://rbbd.febab.org.br/rbbd/article/view/1736. Acesso em: 4 maio. 2023

LOPES, E. A contribuição de Saussure. *In*: ____ **Fundamentos da linguística contemporânea**. São Paulo: Cultrix, 1972.

LOPES, E. **Fundamentos da linguística contemporânea**. 9. ed. São Paulo: Cultrix, 1993.

LOPES, l. L. A. S. **Análise do uso das linguagens controlada e livre nas estratégias de busca em bases de dados**. 2000. Dissertação (Mestrado em Ciência da Informação) - Universidade de Brasília, Brasília, 2000. Disponível em: https://repositorio.unb.br/handle/10482/30528. Acesso em: 10 ago. 2022.

MACULAN, B. C. M. S.; AGANETT, E. C. Desambiguação de relações em tesauros e o seu reúso em ontologias. **Ciência da Informação**, Brasília, DF, v.46 n.1, p.102-119, jan./abr. 2017. Disponível em: https://revista.ibict.br/ciinf/article/view/4017. Acesso em: 03 maio. 2023.

MAIMONE, G. D.; KOBASHI, N. Y.; MOTA, D. Indexação: teoria e métodos. *In:* SILVA, J. F. M.; PALETTA, F. C. (org.). **Tópicos para o ensino de**

biblioteconomia: volume I. São Paulo: ECA/USP, 2016. p. 73-85. Disponível em: https://www.researchgate.net/publication/311707328_Topicos_para_Ensino_de_Bibli oteconomia_V1. Acesso em: 29 ago. 2022.

MARTINS, J. S.; LENZ, M. L.; SILVA, M. B. F.; OLIVEIRA, R. A.; PICHETTI, R. F.; MARIANO, D. C. B.; MARTINS, J. V.; RODRIGUES, S. M. A. F.; BEZERRA, W. R. **Processamentos de Linguagem Natural**. Porto Alegre: Grupo A, 2020. E-book. ISBN 9786556900575. Disponível em: https://integrada.minhabiblioteca.com.br/#/books/9786556900575/. Acesso em: 05 out. 2022.

MATOS, F. F.; MAGALHÃES, L. H.; SOUZA, R. R. Recuperação e classificação de sentimentos de usuários do Twitter em período eleitoral. **Informação & Informação**, [S. l.], v. 25, n. 1, p. 92–114, 2020. DOI: 10.5433/1981-8920.2020v25n1p92. Disponível em: https://ojs.uel.br/revistas/uel/index.php/informacao/article/view/35310. Acesso em: 20 mai. 2023.

MATOS, D. R. S.; OLIVEIRA, F. K. Análise com assistentes virtuais inteligentes: Um estudo de caso com o Google Assistente. **RENOTE**, Porto Alegre, v. 19, n. 1, p. 473–482, 2021. DOI: 10.22456/1679-1916.118537. Disponível em: https://www.seer.ufrgs.br/index.php/renote/article/view/118537. Acesso em: 15 set. 2022.

MEIRELES, M. R. G.; CENDÓN, B. V.; ALMEIDA, P. E. M. Comparação do processo de categorização de documentos utilizando palavras-chave e citações em um domínio de conhecimento restrito. **Transinformação**, v. 28, n. 1, p. 87-96, 2016. DOI: 10.1590/2318-08892016002800007 Acesso em: 02 maio 2023.

MENDONÇA, E. S. A linguística e a ciência da informação: estudos de uma interseção. **Ciência da Informação,** v. 29, n. 3, 2000. DOI: 10.18225/ci.inf.v29i3.873 Acesso em: 12 ago. 2022.

NASCIMENTO, G. D.; MARTINS, G. K.; ALBUQUERQUE, M. E. B. C. Automação da indexação: evidências e tendências da produção científica indexada na Brapci. **Encontros Bibli: Revista Eletrônica de Biblioteconomia e Ciência da Informação**, v. 28, p. 1-20, 2023. DOI: 10.5007/1518-2924.2023.e91956 Acesso em: 28 abr. 2023

NASCIMENTO, S. S. **Introdução à linguística**. 1. ed. Santa Maria, RS: UFSM, NTE, UAB, 2011. Disponível em: https://repositorio.ufsm.br/bitstream/handle/1/17105/Curso_Let-Esp-Lit_Introducao-Linguistica.pdf?sequence=1&isAllowed=y. Acesso em 26 ago. 2022.

NEVES, P. I. CORRÊA, D. A. CAVALCANTI. M. C. Uma análise sobre abordagens e ferramentas para extração de informação. **Revista Militar de Ciência e Tecnologia**. v. 30. 3º tri. 2013. Disponível em: https://rmct.ime.eb.br/arquivos/RMCT_3_tri_2013/RMCT_123_E8A_13.pdf. Acesso em: 08 set. 2022.

OLIVEIRA, F. A. D. **Processamento de Linguagem Natural**: princípios básicos

e a implementação de um Analisador Sintático de Sentenças da Língua Portuguesa. Instituto de Informática da Universidade Federal do Rio Grande do Sul,1999.

OLIVEIRA, F. P.; DIAS, T. M. R.; PINTO, A. L. Modelagem semântica de dados abertos: a viabilidade de aplicação de word embeddings sobre o currículo lattes. **Ciência da Informação**, v. 48, n. 3, 2019. Disponível em: http://hdl.handle.net/20.500.11959/brapci/136520. Acesso em: 02 maio 2023.

PEREIRA, R.; WOLSKI, L. Z.; GONÇALVES, A. L.; CUNHA, C. J. C. A. Mineração de texto e clusterização em estudos bibliométricos: o mapeamento científico de teses e dissertações de um Programa de Pós-Graduação. **Anais do Congresso Internacional de Conhecimento e Inovação – ciki**, [S. l.], v. 1, n. 1, 2022. DOI: 10.48090/ciki.v1i1.1036. Disponível em: https://proceeding.ciki.ufsc.br/index.php/ciki/article/view/1036. Acesso em: 4 maio. 2023.

PINTO, S. C. S. **Processamento de Linguagem Natural e Extração de Conhecimento**. 2015. Dissertação (Mestrado em Engenharia Informática) - Universidade de Coimbra, Coimbra, 2015. Disponível em: http://hdl.handle.net/10316/35676. Acesso em: 23 jul. 2022

PONTES, A. L. Terminologia científica: o que é e como se faz. **Revista de Letras**, v. 1, n. 19, 1997. Disponível em: http://www.periodicos.ufc.br/revletras/article/view/2090. Acesso em: 22 set. 2022.

PRIVATTO, P. I. M. **Uma abordagem para reconhecimento de entidades nomeadas usando conhecimento externo**. 2020. 87 f. Dissertação (Mestrado em Ciência da Computação) - Universidade Estadual Paulista, Rio Claro, 2020. Disponível em: http://hdl.handle.net/11449/194224. Acesso em: 22 set. 2022.

RASTIER, F. Le terme: entre ontologie et linguistique. **La banque des mots**, n.7, 1995. p.35-65. Disponível em: http://www.revue-texto.net/Inedits/Rastier/Rastier_Terme.html. Acesso em: 21 ago. 2022.

ROBREDO, J. **Da Ciência da informação revisitada aos sistemas humanos de informação**. Brasília: Thesaurus, 2003. 262 p

ROSA, J. L. G. **Fundamentos da inteligência artificial**. Rio de Janeiro: LTC, 2011. Disponível em: http://walderson.com/2011-2/IA/FIA.pdf. Acesso em: 23 jul. 2022.

SANTOS, Â. F. S. **Sumarização automática de texto.** 2012. 60 f. Dissertação (Mestrado em Engenharia Informática) - Universidade da Beira Interior, Covilhã, 2012. Disponível em: https://ubibliorum.ubi.pt/bitstream/10400.6/3738/1/Disserta%C3%A7%C3%A3o_M4189.pdf. Acesso em: 02 ago. 2022.

SANTOS, F. M. Análise de conteúdo: a visão de Laurence Bardin. **Revista Eletrônica de Educação**, [S. l.], v. 6, n. 1, p. 383–387, 2012. Disponível em: https://www.reveduc.ufscar.br/index.php/reveduc/article/view/291. Acesso em: 12 ago. 2022.

SANTOS, D. S.; NOGUEIRA, I. C. A.; CARVALHO, C. I. C. Sistema automático de transcrição fonológica para o português. **Texto Livre**, Belo Horizonte, v. 11, n. 2, p. 50–67, 2018. DOI: 10.17851/1983-3652.11.2.50-67. Disponível em: https://periodicos.ufmg.br/index.php/textolivre/article/view/16792. Acesso em: 15 set. 2022.

SANTOS, M. H. **Introdução à inteligência artificial**. Londrina: Editora Saraiva, 2021. Disponível em: https://integrada.minhabiblioteca.com.br/#/books/9786559031245/. Acesso em: 23 jul. 2022.

SARACEVIC, T. Ciência da informação: origem, evolução e relações. **Perspectiva em Ciência da Informação**, Belo Horizonte, v. 1, n. 1, p. 41–62, 1996. Disponível em: https://www.brapci.inf.br/_repositorio/2017/07/pdf_7810a51cca_0000015436.pdf. Acesso em: 22 jul. 2022.

SHEN, M.; LIU, D. R.; HUANG, Y. S. Extracting semantic relations to enrich domain ontologies. **Journal of Intelligent Information Systems**, v. 39, n. 3, p. 749-761, jun. 2012. Disponível em: http://link.springer.com/10.1007/s10844-012-0210-y. Acesso em: 02 maio 2023.

SILVA, B. C. D. DA. O estudo Linguístico-Computacional da Linguagem. **Letras de Hoje**, v. 41, n. 2, 22 set. 2006. Disponível em: https://revistaseletronicas.pucrs.br/ojs/index.php/fale/article/view/597. Acesso em 15 sei. 2022.

SILVA, D. F. **Estudo de funções de similaridade semântica de termos aplicados a um domínio**. 2008. 45 f. Monografia (Graduação em Ciência da Computação) - Universidade Federal de Pernambuco, Recife, 2008. Disponível em: https://www.cin.ufpe.br/~tg/2007-2/dfs3.pdf. Acesso em: 02 mai. 2023.

SILVA, E. L.; MENEZES, E. M. **Metodologia da pesquisa e elaboração de dissertação**. 4. ed. rev. atual. Florianópolis: UFSC, 2005. Disponível em: https://cursos.unipampa.edu.br/cursos/ppgcb/files/2011/03/Metodologia-da-Pesquisa-3a-edicao.pdf. Acesso em: 02 ago. 2022.

SOUZA, A. D.; FELIPE, E. R. Processamento de Linguagem Natural aplicado à anamneses do domínio da ginecologia. **Fronteiras da Representação do Conhecimento**, [S. l.], v. 1, n. 2, p. 51–69, 2021. Disponível em: https://periodicos.ufmg.br/index.php/fronteiras-rc/article/view/37308. Acesso em: 5 out. 2022.

SOUZA, A. R.; SCHIRRU, L.; ALVARENGA, M. B. Direitos autorais e mineração de dados e textos no combate à covid-19 no brasil. **Liinc em revista**, v. 16, 2020. DOI: 10.18617/liinc.v16i2.5536 Acesso em: 04 maio 2023

SOUZA, C. L. **Abordagem Computacional para Criação de Neologismos Terminológicos em Línguas de Sinais**. 2018. Tese (Doutorado em Modelagem Matemática e Computacional) – Centro Federal de Educação Tecnológica de Minas

Gerais, Belo Horizontes, 2018. Disponível em: https://www.researchgate.net/publication/326302826_Abordagem_Computacional_para_Criacao_de_Neologismos_Terminologicos_em_Linguas_de_Sinais/citations#fullTextFileContent. Acesso em: 28 ago. 2022.

SOUZA, M.; SOUZA, R. R. Modelagem de tópicos: resumir e organizar corpus de dados por meio de algoritmos de aprendizagem de máquina. **Múltiplos Olhares em Ciência da Informação**, v. 9 No. 2, n. 2, 2019. Disponível em: http://hdl.handle.net/20.500.11959/brapci/137081. Acesso em: 01 maio 2023.

SOUZA, O. de; TABOSA, H. R.; OLIVEIRA, D. M. de; OLIVEIRA, M. H. de S. Um método de sumarização automática de textos através de dados estatísticos e Processamento de Linguagem Natural. **Informação & Sociedade: Estudos**, [S. l.], v. 27, n. 3, 2017. Disponível em: https://periodicos.ufpb.br/ojs2/index.php/ies/article/view/32571. Acesso em: 1 maio. 2023.

SOUZA, R. R. **Uma proposta para metodologia para escolha automática de descritores utilizando sintagmas nominais**. 2005. Tese (Doutorado em Ciência da Informação) - Universidade Federal de Minas Gerais, Belo Horizonte, 2005. Disponível em: http://hdl.handle.net/1843/RRSA-6GGGUF. Acesso em: 12 ago. 2022.

TEMMERMMAN, R. Sociocognitive terminology theory. In: **Simpósio Internacional de Verano de Terminología**: Terminologia y cognición. Barcelona: IULA-UPF, 2001. Disponível em: https://euralex.org/elx_proceedings/Euralex2000/053_Rita%20TEMMERMANN_Training%20Terminographers_the%20Sociocognitive%20Approach.pdf. Acesso em: 22 ago. 2022.

VALE, E. A. Linguagens de indexação. In: SMIT, J. W. (org.). **Análise documentária**: a análise da síntese. Brasília: Ibict, 1987. 133 p. Disponível em: http://livroaberto.ibict.br/handle/1/101 Acesso em: 27 set. 2022.

WAZLAWICK, R. S. **Metodologia de pesquisa para Ciência da Computação**. Rio de Janeiro: Grupo GEN, 2020. E-book. ISBN 9788595157712. Disponível em: https://integrada.minhabiblioteca.com.br/#/books/9788595157712/. Acesso em: 12 ago. 2022.

WIVES, L. K.; LOH, S. Tecnologias de descoberta de conhecimento em informações textuais; ênfase em agrupamento de informações. In: OFICINA DE INTELIGÊNCIA ARTIFICIAL (OIA) III, 1999, Pelotas (RS). **Proceedings…** Pelotas: EDUCAT, 1999. p. 28-48.

APÊNDICES

APÊNDICE A - Artigos selecionados na categoria Fundamentação e conceituais

REFERÊNCIAS	BASES DE DADOS
ALVES, R. P. D. S.; CURTY, R. G.; TREVISAN, G. L. Análise da produção científica do periódico JASIS&T sob a ótica dos três paradigmas da ciência da informação. **Tendências da Pesquisa Brasileira em Ciência da Informação**, v. 11, n. 1, 2018. Disponível em: https://brapci.inf.br/index.php/res/v/151747. Acesso em: 22 jul. 2022.	BRAPCI
FALCÃO, L. C. J.; LOPES, B.; SOUZA, R. R. Absorção das Tarefas de Processamento de Linguagem Natural (NLP) pela Ciência da Informação (CI): uma revisão da literatura para tangibilização do uso de NLP pela CI. **Em Questão,** n. online, 2021. Disponível em: https://brapci.inf.br/index.php/res/v/164766. Acesso em: 22 jul. 2022.	BRAPCI
GOMES D. C.; OLIVEIRA, L. E. S.; CUBAS, M. R.; BARRA, C. M. Uso de ferramentas computacionais como auxílio ao método de mapeamento cruzado entre terminologias clínicas. **Texto Contexto Enfermagem**, v. 28, 2019. Disponível em: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85066810944&doi=10.1590%2f1980-265x-tce-2017-0187&partnerID=40 DOI: 10.1590/1980-265x-tce-2017-0187. Acesso em: 22 jul. 2022.	SCOPUS
LEWINSKI, N. A.; MCINNES, B. T. Using natural language processing techniques to inform research on nanotechnology. **Beilstein Journal of Nanotechnology**, v. 6, n. 1, 2015, p. 1439-1449. Disponível em: https://www.scopus.com/inward/record.uri?eid=2-s2.0-84947943260&doi=10.3762%2fbjnano.6.149&partnerID=40&md5=5de12 DOI: 10.3762/bjnano.6.149. Acesso em: 22 jul. 2022.	SCOPUS
MAIA, E. H. B.; BAX, M. P. Um estudo bibliográfico sobre ligação de entidades. **Informação & Informação**, n. 2, v. 21, p. 245-291, 2016. Disponível em: https://brapci.inf.br/index.php/res/v/33777. Acesso em: 22 jul. 2022.	BRAPCI
Metodologias, ferramentas e aplicações da inteligência artificial nas diferentes linhas do combate a Covid-19. **Revista Folha de Rosto**, n. 2, v. 6, p. 44-57, 2020. Disponível em: https://brapci.inf.br/index.php/res/v/145722. Acesso em: 22 jul. 2022.	BRAPCI
PUERTA-DíAZ, M.; MIRA, B. S.; MARTÍNEZ-ÁVILA, D.; OVALLE-PERANDONES, M.; GRÁCIO, M. C. C. O Processamento de Linguagem Natural nos Estudos Métricos da Informação: uma análise dos artigos indexados pela Web of Science (2000- 2019). **Encontros Bibli: Revista Eletrônica de Biblioteconomia e Ciência da Informação**, v. 26, p. 1-24, 2021. Disponível em: <https://brapci.inf.br/index.php/res/v/156831>. Acesso em: 22 jul. 2022.	BRAPCI
SILVA, S. R B. CORRÊA, R. F. Sistemas de Indexação automática por atribuição: uma análise comparativa. **Encontros Bibli: Revista Eletrônica de Biblioteconomia e Ciência da Informação**, v. 25, p. 1-25, 2020. Disponível em: <https://brapci.inf.br/index.php/res/v/142221>. Acesso em: 22 jul. 2022.	BRAPCI
SOUZA, O.; TABOSA, H. R. Estudo sobre contribuição da Ciência da Informação em pesquisas sobre Tecnologias Assistivas. **Comunicação & Informação,** n. 1, v. 21, p. 70-88, 2018. Disponível em: <https://brapci.inf.br/index.php/res/v/67501>. Acesso em: 22-jul.-2022.	BRAPCI

APÊNDICE B - Artigos selecionados na categoria Pré-processamento de texto

REFERÊNCIAS	BASE DE DADOS
TABOSA, H. R.; SOUZA, O.; CÂNDIDO, J. C. D. S.; MELO, A. C. A. U.; REIS, K. G. B. Avaliação do desempenho de um software de sumarização automática de textos. **Informação & Informação**, n. 1, v. 25, p. 189-210, 2020. Disponível em: <https://brapci.inf.br/index.php/res/v/137794>. Acesso em: 22 jul. 2022.	BRAPCI
GOLDSTEIN, A.; SHAHAR, Y. An automated knowledge-based textual summarization system for longitudinal, multivariate clinical data. **Journal of Biomedical Informatics**, 61, 2016. p. 159-175. Disponível em: https://www.scopus.com/inward/record.uri?eid=2-s2.0-84962704669&doi=10.1016%2fj.jbi.2016.03.022&partnerID=40&md5=f71 DOI: 10.1016/j.jbi.2016.03.022. Acesso em: 22 jul. 2022.	SCOPUS
MORITA, H.; SAKAI, T.; OKUMURA, M. Query Snowball: A Co-occurrence-based Approach to Multi-document Summarization for Question Answering. **IPSJ Online Transactions**, v. 5, 2012, p. 124-129. Disponível em: https://www.scopus.com/inward/record.uri?eid=2-s2.0-84937045251&doi=10.2197%2fipsjtrans.5.124&partnerID=40&md5=f447 DOI: 10.2197/ipsjtrans.5.124. Acesso em: 22 jul. 2022.	SCOPUS
BUI, D. D. A., WYATT, M., CIMINO, J. J. The UAB Informatics Institute and 2016 CEGS N-GRID de-identification shared task challenge. **Journal of Biomedical Informatics**, v. 75, 2017. p. S54-S61. Disponível em: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85018758241&doi=10.1016%2fj.jbi.2017.05.001&partnerID=40&md5=40 DOI: 10.1016/j.jbi.2017.05.001. Acesso em: 22 jul. 2022. SCOPUS	SCOPUS
MARTIM, H.; LIMA, J. A. O.; ARAUJO, L. C. Base de normas jurídicas brasileiras: uma iniciativa de open government data. **Perspectivas em Ciência da Informação**, v. 23, n. 4, p. 133-149, 2018. Disponível em: http://hdl.handle.net/20.500.11959/brapci/108424. Acesso em: 22 jul. 2022	BRAPCI

APÊNDICE C - Artigos selecionados na categoria Análise semântica e Representação de texto

REFERÊNCIAS	BASES DE DADOS
BALLATORE, A.; BERTOLOTTO, M.; WILSON, D. C. An Evaluative Baseline for Geo-Semantic Relatedness and Similarity. **GeoInformatica**, v. 18, n. 4, 2014, p. 747-767. Disponível em: https://www.scopus.com/inward/record.uri?eid=2-s2.0-84908118044&doi=10.1007%2fs10707-013-0197-8&partnerID=40&md5= DOI: 10.1007/s10707-013-0197-8. Acesso em: 22 jul. 2022.	SCOPUS
HARISPE, S., SÁNCHEZ, D., RANWEZ, S., JANAQI, S., MONTMAIN, J. A framework for unifying ontology-based semantic similarity measures: A study in the biomedical domai. **Journal of Biomedical Informatics**, v. 48, 2014, p. 38-53. Disponível em: https://www.scopus.com/inward/record.uri?eid=2-s2.0-84899492201&doi=10.1016%2fj.jbi.2013.11.006&partnerID=40&md5=e7 DOI: 10.1016/j.jbi.2013.11.006. Acesso em: 22 jul. 2022.	SCOPUS
MARTINEZ-GIL, J. CoTO: A novel approach for fuzzy aggregation of semantic similarity measures. **Cognitive Systems Research**, v. 40, 2016, p. 8-17. Disponível em: https://www.scopus.com/inward/record.uri?eid=2-s2.0-84981731334&doi=10.1016%2fj.cogsys.2016.01.001&partnerID=40&md DOI: 10.1016/j.cogsys.2016.01.001. Acesso em: 22 jul. 2022.	SCOPUS
SOLEIMANDARABI, M. N.; MIRROSHANDEL, S. A. A Novel Approach for Computing Semantic Relatedness of Geographic Terms. **Indian Journal of Science and Technology**, v. 8, n. 27, 2015. Disponível em: https://www.scopus.com/inward/record.uri?eid=2-s2.0-84979747624&doi=10.17485%2fijst%2f2015%2fv8i27%2f60811&partner DOI: 10.17485/ijst/2015/v8i27/60811. Acesso em: 22 jul. 2022.	SCOPUS
RUBIN, V. L., CHEN, Y., CONROY, N. J. Deception detection for news: Three types of fakes. **Proceedings of the Association for Information Science and Technology**, v. 52, n. 1. 2015. p. 1-4. Disponível em: https://www.scopus.com/inward/record.uri?eid=2-s2.0-84987753820&doi=10.1002%2fpra2.2015.145052010083&partnerID=40 DOI: 10.1002/pra2.2015.145052010083. Acesso em: 22 jul. 2022.	SCOPUS
RUBIN, V.L., CONROY, N. Discerning truth from deception: Human judgments and automation efforts. **First Monday**, v. 17, n. 3, 2012. Disponível em: https://www.scopus.com/inward/record.uri?eid=2-s2.0-84864013036&doi=10.5210%2ffm.v17i3.3933&partnerID=40&md5=5c5e DOI: 10.5210/fm.v17i3.3933. Acesso em: 22 jul. 2022.	SCOPUS
RUBIN, V.L., CHEN, Y. Information manipulation classification theory for LIS and NLP. **Proceedings of the ASIST Annual Meeting**, v. 49, n. 1, 2012. p. 1-5. Disponível em: https://www.scopus.com/inward/record.uri?eid=2-s2.0-84878536551&doi=10.1002%2fmeet.14504901353&partnerID=40&md5= DOI: 10.1002/meet.14504901353. Acesso em: 22 jul. 2022.	SCOPUS
SANTOS, J.; ANDRADE, F.; JORGE, E. M. F.; BATISTA, J.; SABA, H. Redes complexas de homônimos para análise semântica textual. **Informação & Informação**, v. 22, n. 1, p. 293-305, 2017. DOI: 10.5433/1981-8920.2017v22n1p293 Acesso em: 22 jul. 2022.	BRAPCI
CONEGLIAN, C. S.; SANTARÉM SEGUNDO, J. E. Inteligência artificial e ferramentas da web semântica aplicadas a recuperação da informação: um modelo conceitual com foco na linguagem natural. **Informação & Informação**, v. 27, n. 1, p. 625-651, 2022. DOI: 10.5433/1981-8920.2022v27n1p625 Acesso em: 22 jul. 2022.	BRAPCI

APÊNDICE D - Artigos selecionados na categoria de Extração de Informação e Mineração de texto

REFERÊNCIAS	BASES DE DADOS
TEIXEIRA, W. R.; LIMA, J. A. O.; ARAUJO, L. C.; VIERO, D. M.; SANTANA, F. F.; HERINGER, F. R. A.; MARTIM, H.; VIEIRA FILHO, J. J. Exemplo de extração de definições em textos articulados de normas jurídicas com o apoio do Processamento de Linguagem Natural. **CAJUR - Caderno de Informações Jurídicas,** v. 6, n. 1, 2019. Disponível em: http://hdl.handle.net/20.500.11959/brapci/119039. Acesso em: 22 jul. 2022.	BRAPCI
SILVA, E. M.; SOUZA, R. R. Fundamentos em processamento de linguagem natural: uma proposta para extração de bigramas. **Encontros Bibli: Revista Eletrônica de Biblioteconomia e Ciência da Informação**, v. 19, n. 40, p. 1-32, 2014. DOI: 10.5007/1518-2924.2014v19n40p1 Acesso em: 22 jul. 2022.	BRAPCI
SOUZA, A. D.; FELIPE, E. R. Processamento de linguagem natural aplicado à anamneses do domínio da ginecologia**. Fronteiras da Representação do Conhecimento**, v. 1, p. 51-69, 2021. Disponível em: http://hdl.handle.net/20.500.11959/brapci/194014. Acesso em: 22 jul. 2022.	BRAPCI
BEJAN, C. A., WEI, W.-Q., DENNY, J. C. Assessing the role of a medication-indication resource in the treatment relation extraction from clinical text. **Journal of the American Medical Informatics Association**, v. 22, 2014, pp. e162-e176. Disponível em: https://www.scopus.com/inward/record.uri?eid=2-s2.0-84986917859&doi=10.1136%2famiajnl-2014-002954&partnerID=40&md DOI: 10.1136/amiajnl-2014-002954. Acesso em: 22 jul. 2022.	SCOPUS
BOZKURT, S., LIPSON, J. A., SENOL, U., RUBIN, D. L. Automatic abstraction of imaging observations with their characteristics from mammography reports. **Journal of the American Medical Informatics Association**, v. 22, 2014. p. e81-e92. Disponível em: https://www.scopus.com/inward/record.uri?eid=2-s2.0-84958695164&doi=10.1136%2famiajnl-2014-003009&partnerID=40&md DOI: 10.1136/amiajnl-2014-003009. Acesso em: 22 jul. 2022.	SCOPUS
YANG, X., BIAN, J., HOGAN, W.R., WU, Y. Clinical Relation Extraction Using Transformer-based Models. **Journal of the American Medical Informatics Association,** v. 27, n. 12, 2020. p. 1935-1942. Disponível em: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85098465042&doi=10.1093%2fjamia%2focaa189&partnerID=40&md5=4 DOI: 10.1093/jamia/ocaa189. Acesso em: 22 jul. 2022.	SCOPUS
JACKSON, R., KARTOGLU, I., STRINGER, C., GORRELL, G., ROBERTS, A., SONG, X., WU, H., AGRAWAL, A., LUI, K., GROZA, T., LEWSLEY, D., NORTHWOOD, D., FOLARIN, A., STEWART, R., DOBSON, R. CogStack - Experiences of deploying integrated information retrieval and extraction services in a large National Health Service Foundation Trust hospital. **BMC Medical Informatics and Decision Making,** v. 18, n.1, 2018. Disponível em: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85049252790&doi=10.1186%2fs12911-018-0623-9&partnerID=40&md5= DOI: 10.1186/s12911-018-0623-9. Acesso em: 22 jul. 2022.	SCOPUS
SHI, J., ZHENG, M., YAO, L., GE, Y. Developing a healthcare dataset information resource (DIR) based on Semantic Web. **BMC Medical Genomics**, v. 11, 2018. Disponível em: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85056706345&doi=10.1186%2fs12920-018-0411-5&partnerID=40&md5= DOI: 10.1186/s12920-018-0411-5. Acesso em: 22 jul. 2022.	SCOPUS
BREITENSTEIN, M.K., LIU, H., MAXWELL, K.N., PATHAK, J., ZHANG, R. Electronic Health Record Phenotypes for Precision Medicine: Perspectives and Caveats from Treatment of Breast Cancer at a Single Institution. **Clinical and**	SCOPUS

Translational Science, v. 11, n. 1, 2018. p. 85-92. Disponível em: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85040248565&doi=10.1111%2fcts.12514&partnerID=40&md5=97fcf39b9 DOI: 10.1111/cts.12514. Acesso em: 22 jul. 2022.	
KANG, T., ZHANG, S., TANG, Y., HRUBY, W., RUSANOV, A., ELHADAD, N., WENG, C. EliIE: An open-source information extraction system for clinical trial eligibility criteria. **Journal of the American Medical Informatics Association**, v. 24, n. 6, 2017. Disponível em: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85032911356&doi=10.1093%2fjamia%2focx019&partnerID=40&md5=7a DOI: 10.1093/jamia/ocx019. Acesso em: 22 jul. 2022.	SCOPUS
DAS, R. D., PURVES, R. S. Exploring the Potential of Twitter to Understand Traffic Events and Their Locations in Greater Mumbai, India. **IEEE Transactions on Intelligent Transportation Systems**, v. 21, n. 12, 2020. p. 5213-5222. Disponível em: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85096166799&doi=10.1109%2fTITS.2019.2950782&partnerID=40&md5 DOI: 10.1109/TITS.2019.2950782. Acesso em: 22 jul. 2022.	SCOPUS
TREUDE, C., ROBILLARD, M. P., DAGENAIS, B. Extracting development tasks to navigate software documentation. **IEEE Transactions on Software Engineering**, v. 41, n. 6, 2015. Disponível em: https://www.scopus.com/inward/record.uri?eid=2-s2.0-84933054135&doi=10.1109%2fTSE.2014.2387172&partnerID=40&md5= DOI: 10.1109/TSE.2014.2387172. Acesso: 22 jul. 2022.	SCOPUS
JIANG, G., DHRUVA, S.S., CHEN, J., SCHULZ, W.L., DOSHI, A.A., NOSEWORTHY, P.A., ZHANG, S., YU, Y., PATRICK YOUNG, H., BRANDT, E., ERVIN, K.R., SHAH, N.D., ROSS, J.S., COPLAN, P., DROZDA, J.P. Feasibility of capturing real-world data from health information technology systems at multiple centers to assess cardiac ablation device outcomes: A fit-for-purpose informatics analysis report. **Journal of the American Medical Informatics Association: JAMIA**, v. 28, n. 10, 2021. p. 2241-2250. Disponível em: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85116959104&doi=10.1093%2fjamia%2focab117&partnerID=40&md5=2 DOI: 10.1093/jamia/ocab117. Acesso em: 22 jul. 2022.	SCOPUS
BOYTCHEVA, S., ANGELOVA, G., ANGELOV, Z., TCHARAKTCHIEV, D. Text mining and big data analytics for retrospective analysis of clinical texts from outpatient care **Cybernetics and Information Technologies**, v. 15, n. 4, 2015. p. 58-77. Disponível em: https://www.scopus.com/inward/record.uri?eid=2-s2.0-84959246070&doi=10.1515%2fcait-2015-0055&partnerID=40&md5=467 DOI: 10.1515/cait-2015-0055. Acesso em: 22 jul. 2022	SCOPUS
BANERJI, A., LAI, K. H., LI, Y., SAFF, R. R., CAMARGO, C. A., BLUMENTHAL, K. G., ZHOU, L. Natural Language Processing Combined with ICD-9-CM Codes as a Novel Method to Study the Epidemiology of Allergic Drug Reactions. **Journal of Allergy and Clinical Immunology: In Practice**, v. 8, n. 3, 2020. p. 1032-1038. Disponível em: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85077694717&doi=10.1016%2fj.jaip.2019.12.007&partnerID=40&md5=5 DOI: 10.1016/j.jaip.2019.12.007. Acesso em: 22 jul. 2022.	SCOPUS
LIN, C., DLIGACH, D., MILLER, T.A., BETHARD, S., SAVOVA, G.K. Multilayered temporal modeling for the clinical domain. **Journal of the American Medical Informatics Association,** v. 23, n. 2, 2016. p. 387-395. Disponível em: https://www.scopus.com/inward/record.uri?eid=2-s2.0-84963741141&doi=10.1093%2fjamia%2focv113&partnerID=40&md5=b0 DOI: 10.1093/jamia/ocv113. Acesso em: 22 jul. 2022.	SCOPUS
DU, J., CHEN, T., ZHANG, L. Measuring the interactions between health demand, informatics supply, and technological applications in digital medical	SCOPUS

innovation for China: Content mapping and analysis **JMIR Medical Informatics**, v. 9, n. 7, 2021. Disponível em: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85110026316&doi=10.2196%2f26393&partnerID=40&md5=eae9d6ff736 DOI: 10.2196/26393. Acesso em: 22 jul. 2022	
MARRONE, M. Application of entity linking to identify research fronts and trends. **Scientometrics**, v. 122, n. 1, 2020. p. 357-379. Disponível em: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85074816789&doi=10.1007%2fs11192-019-03274-x&partnerID=40&md5 DOI: 10.1007/s11192-019-03274-x. Acesso em: 22 jul. 2022.	SCOPUS
MESQUITA, L. A. L.; DIAS, C. C.; SOUZA, R. R. O fluxo temporal de termos relevantes: uma análise em teses da UFMG de 2007 a 2018 nas Ciências Humanas. **Múltiplos Olhares em Ciência da Informação**, n. forped-ppggoc - 2021, 2021. Disponível em: http://hdl.handle.net/20.500.11959/brapci/171061. Acesso em: 22 jul. 2022.	BRAPCI
SCHIESSL, M.; BRÄSCHER, M. Descoberta de Conhecimento em Texto aplicada a um Sistema de Atendimento ao Consumidor. **Revista Ibero-Americana de Ciência da Informação**, n. 2, v. 4 No 2, p. 94-111, 2011. Disponível em: <https://brapci.inf.br/index.php/res/v/72887>. Acesso em: 22 jul. 2022.	BRAPCI
LI, X., LI, J., WU, Y. A global optimization approach to multi-polarity sentiment analysis. **PLoS ONE**, v. 10, n. 4, 2015. Disponível em: https://www.scopus.com/inward/record.uri?eid=2-s2.0-84929485519&doi=10.1371%2fjournal.pone.0124672&partnerID=40&md DOI: 10.1371/journal.pone.0124672. Acesso em: 22 jul. 2022.	SCOPUS
TOURASSI, G., YOON, H.-J., XU, S. A novel web informatics approach for automated surveillance of cancer mortality trends. **Journal of Biomedical Informatics**, v. 61, 2016. p. 110-118. Disponível em: https://www.scopus.com/inward/record.uri?eid=2-s2.0-84962701015&doi=10.1016%2fj.jbi.2016.03.027&partnerID=40&md5=1e DOI: 10.1016/j.jbi.2016.03.027. Acesso em: 22 jul. 2022.	SCOPUS
SHAH, R. U., MUKHERJEE, R., ZHANG, Y., JONES, A. E., SPRINGER, J., HACKETT, I., STEINBERG, B. A., LLOYD-JONES, D. M., Chapman, W. W Impact of different electronic cohort definitions to identify patients with atrial fibrillation from the electronic medical record. **Journal of the American Heart Association**, v. 9, n. 5, 2020. Disponível em: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85090263321&doi=10.1161%2fJAHA.119.014527&partnerID=40&md5=c DOI: 10.1161/JAHA.119.014527. Acesso em: 22 jul. 2022.	SCOPUS
MIRANKER, M., GIORDANO, A. Text mining and semantic triples: Spatial analyses of text in applied humanitarian forensic research. **Digital Geography and Society**, v. 1, 2020. Disponível em: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85125606663&doi=10.1016%2fj.diggeo.2020.100005&partnerID=40&md DOI: 10.1016/j.diggeo.2020.100005. Acesso em: 22 jul. 2022.	SCOPUS
HULUBA, A.-M., KINGDON, J., MCLAREN, I. The UK Online Gender Audit 2018: A comprehensive audit of gender within the UK's online environment. **Heliyon**, v. 4, n. 12, 2018. Disponível em: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85057759135&doi=10.1016%2fj.heliyon.2018.e01001&partnerID=40&md DOI: 10.1016/j.heliyon.2018.e01001. Acesso em: 22 jul. 2022.	SCOPUS
Lei, Y., Xu, S., Zhou, L. User Behaviors and User-Generated Content in Chinese Online Health Communities: Comparative Study. **Journal of Medical Internet Research**, v. 23, n. 12, 2021. Disponível em:	SCOPUS

https://www.scopus.com/inward/record.uri?eid=2-s2.0-85121983303&doi=10.2196%2f19183&partnerID=40&md5=7613f2864d3 DOI: 10.2196/19183. Acesso em: 22 jul. 2020.	
CERNILE, G., HERITAGE, T., SEBIRE, N.J., GORDON, B., SCHWERING, T., KAZEMLOU, S., BORECKI, Y. Network graph representation of COVID-19 scientific publications to aid knowledge discovery BMJ **Health and Care Informatics,** v. 28, n. 1, 2021. Disponível em: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85099197229&doi=10.1136%2fbmjhci-2020-100254&partnerID=40&md5 DOI: 10.1136/bmjhci-2020-100254. Acesso em: 22 jul. 2020.	SCOPUS
SÁNCHEZ, D., BATET, M., VIEJO, A. Utility-preserving privacy protection of textual healthcare documents**. Journal of Biomedical Informatics**, v. 52, 2014. p. 189-198. Disponível em: https://www.scopus.com/inward/record.uri?eid=2-s2.0-84919844382&doi=10.1016%2fj.jbi.2014.06.008&partnerID=40&md5=8fd DOI: 10.1016/j.jbi.2014.06.008. Acesso em: 22 jul. 2022.	SCOPUS
MAGALHÃES, L. H.; SOUZA, R. R. Agrupamento automático de notícias de jornais on-line usando técnicas de machine learning para clustering de textos no idioma português. **Múltiplos Olhares em Ciência da Informação**, v. 9 No. 2, n. 2, 2019. Disponível em: http://hdl.handle.net/20.500.11959/brapci/137097. Acesso em: 22 jul. 2022.	BRAPCI
LIANG, J., ZHANG, Z., FAN, L., SHEN, D., CHEN, Z., XU, J., GE, F., XIN, J., LEI, J. A Comparison of the development of medical informatics in china and that in western countries from 2008 to 2018: A bibliometric analysis of official journal publications. **Journal of Healthcare Engineering**, 2020. Disponível em: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85094682637&doi=10.1155%2f2020%2f8822311&partnerID=40&md5=6 DOI: 10.1155/2020/8822311. Acesso em: 22 jul. 2022.	SCOPUS
ANDRUSHCHENKO, M., SANDBERG, K., TURUNEN, R., MARJANEN, J., HATAVARA, M., KURUNMÄKI, J., NUMMENMAA, T., HYVÄRINEN, M., TERÄS, K., PELTONEN, J., NUMMENMAA, J. Using parsed and annotated corpora to analyze parliamentarians' talk in Finland. **Journal of the Association for Information Science and Technology**, v. 73, n. 2, 2022. p. 288-302. Disponível em: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85107229944&doi=10.1002%2fasi.24500&partnerID=40&md5=2417bbd6 DOI: 10.1002/asi.24500. Acesso em: 22 jul. 2022.	SCOPUS
HUANG, D., WANG, S., REN, F. Creating Chinese-English comparable corpora**. IEICE Transactions on Information and Systems**, n. 8, 2013. p. 1853-1861. Disponível em: https://www.scopus.com/inward/record.uri?eid=2-s2.0-84882693326&doi=10.1587%2ftransinf.E96.D.1853&partnerID=40&md5 DOI: 10.1587/transinf.E96.D.185319. Acesso em: 22 jul. 2022.	SCOPUS
KOLHAR, M., ALAMEEN, A. University learning and anti-plagiarism back-end services. **Computers, Materials and Continua**, v. 66, n. 2, 2020. p. 1215-1226. Disponível em: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85097127272&doi=10.32604%2fcmc.2020.012658&partnerID=40&md5= DOI: 10.32604/cmc.2020.012658. Acesso em: 22 jul. 2022.	SCOPUS
KOZA, W., FILIPPO, D., COTIK, V., STRICKER, V., MUÑOZ, M., GODOY, N., RIVAS, N., MARTÍNEZ-GAMBOA, R. Automatic Detection of Negated Findings in Radiological Reports for Spanish Language: Methodology Based on Lexicon-Grammatical Information Processing. **Journal of Digital Imaging**, v. 32, n. 1, 2019. p. 19-29. Disponível em: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85051667846&doi=10.1007%2fs10278-018-0113-8&partnerID=40&md5= DOI: 10.1007/s10278-018-0113-8. Acesso em: 22 jul. 2022.	SCOPUS.

APÊNDICE E - Artigos selecionados na categoria Aplicação de Modelagem de tópicos e Classificação de texto

REFERÊNCIAS	BASES DE DADOS
PUERTA-DìAZ, M.; OVALLE-PERANDONES, M.; MARTÍNEZ-ÁVILA, D. Padrões emergentes e tendências da estrutura científica internacional no domínio "discurso do ódio". **Revista Ibero-Americana de Ciência da Informação**, v. 13, p. 963-978, 2020. DOI: 10.26512/rici.v13.n3.2020.33017 Acesso em: 22 jul. 2022.	BRAPCI
WANG, T., ZHOU, Z., HU, X., LIU, Z., DING, Y., CAI, Z. Latent topics resonance in scientific literature and commentaries: evidences from natural language processing approach. **Heliyon**, v. 4, n. 6, 2018. Disponível em: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85048804006&doi=10.1016%2fj.heliyon.2018.e00659&partnerID=40&md. Acesso em: 22 jul. 2022.	SCOPUS
ARNOLD, C.W., OH, A., CHEN, S., SPEIER, W. Evaluating topic model interpretability from a primary care physician perspective. **Computer Methods and Programs in Biomedicine**, v. 124, 2016, p. 67-75. Disponível em: https://www.scopus.com/inward/record.uri?eid=2-s2.0-84954527214&doi=10.1016%2fj.cmpb.2015.10.014&partnerID=40&md5= DOI: 10.1016/j.cmpb.2015.10.014. Acesso em: 22 jul. 2022.	SCOPUS
AMBALAVANAN, A. K., DEVARAKONDA, M. V. Using the contextual language model BERT for multi-criteria classification of scientific articles. **Journal of Biomedical Informatics**, v. 112, 2020. Disponível em: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85093935873&doi=10.1016%2fj.jbi.2020.103578&partnerID=40&md5=6c DOI: 10.1016/j.jbi.2020.103578. Acesso em: 22 jul. 2022.	SCOPUS
GUIMARÃES, L. M. S.; MEIRELES, M. R. G.; ALMEIDA, P. E. M. Avaliação das etapas de pré-processamento e de treinamento em algoritmos de classificação de textos no contexto da recuperação da informação. **Perspectivas em Ciência da Informação**, v. 24, n. 1, p. 169-190, 2019. Disponível em: http://hdl.handle.net/20.500.11959/brapci/112210. Acesso em: 22 JUL. 2022.	BRAPCI
REDD, D., FRECH, T.M., MURTAUGH, M.A., RHIANNON, J., ZENG, Q.T. Informatics can identify systemic sclerosis (SSc) patients at risk for scleroderma renal crisis. **Computers in Biology and Medicine**, v. 53, 2014. p. 203-205. Disponível em: https://www.scopus.com/inward/record.uri?eid=2-s2.0-84906719270&doi=10.1016%2fj.compbiomed.2014.07.022&partnerID=4 DOI: 10.1016/j.compbiomed.2014.07.022. Acesso em: 22 jul. 2022.	SCOPUS
DLIGACH, D., AFSHAR, M., MILLER, T. Toward a clinical text encoder: Pretraining for clinical natural language processing with applications to substance misuse. **Journal of the American Medical Informatics Association**, v. 26, n. 11, 2019. p. 1272-1278. Disponível em: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85073184468&doi=10.1093%2fjamia%2focz072&partnerID=40&md5=8e DOI: 10.1093/jamia/ocz072. Acesso em: 22 jul. 2022.	SCOPUS
PEARCE, C., MCLEOD, A., PATRICK, J., FERRIGI, J., BAINBRIDGE, M.M., RINEHART, N., FRAGKOUDI, A. Coding and classifying GP data: The POLAR project. **BMJ Health and Care Informatics**, v. 26, n. 1, 2019. Disponível em: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85074822562&doi=10.1136%2fbmjhci-2019-100009&partnerID=40&md5 DOI: 10.1136/bmjhci-2019-100009. Acesso em: 22 jul. 2020.	SCOPUS

KUWANA, A., OB A, A., SAWAI, R., PAIK, I. Automatic Taxonomy Classification by Pretrained Language Model**. Electronics (Switzerland),** v. 10, n. 21, 2021. Disponível em: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85118191099&doi=10.3390%2felectronics10212656&partnerID=40&md5 DOI: 10.3390/electronics10212656	SCOPUS
HAN, D., WANG, S., JIANG, C., JIANG, X., KIM, H.-E., SUN, J., OHNO-MACHADO, L. Trends in biomedical informatics: automated topic analysis of JAMIA articles. **Journal of the American Medical Informatics Association**, v. 22, n. 6, 2015. p. 1153-1163. Disponível em: https://www.scopus.com/inward/record.uri?eid=2-s2.0-84954239460&doi=10.1093%2fjamia%2focv157&partnerID=40&md5=165c0ba010c4478dc26e62af0d1d1f8d DOI: 10.1093/jamia/ocv157. Acesso em: 22 jul. 2022.	SCOPUS
MEDEIROS, L. F.; MOSER, A.; SANTOS, N. D. Assistente de conhecimento conceitual como um sistema intencional para processos tutoriais em educação a distância. **Perspectivas em Gestão & Conhecimento**, v. 5, n. 1, p. 155-168, 2015. Disponível em: http://hdl.handle.net/20.500.11959/brapci/49989. Acesso em: 22 jul. 2022.	BRAPCI