UNIVERSIDADE FEDERAL DE SANTA CATARINA
CAMPUS FLORIANÓPOLIS
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA, GESTÃO E MÍDIA DO CONHECIMENTO

Luciano Zamperetti Wolski

Modelo de classificação de patentes baseado em representação vetorial densa, técnicas de ordenação e explicitação do conhecimento

Florianópolis
2023

Luciano Zamperetti Wolski

Modelo de classificação de patentes baseado em representação vetorial densa, técnicas de ordenação e explicitação do conhecimento

Tese submetida ao Programa de Pós-Graduação em Engenharia, Gestão e Mídia do Conhecimento da Universidade Federal de Santa Catarina como requisito para a obtenção do título de Doutor em Engenharia e Gestão do Conhecimento. Área de concentração: Engenharia do Conhecimento. Linha de pesquisa: Teoria e Prática em Engenharia do Conhecimento.

Orientador: Prof. Dr. Alexandre Leopoldo Gonçalves
Coorientador: Prof. Dr. José Leomar Todesco

Florianópolis
2023

Wolski, Luciano Zamperetti
Modelo de Classificação de Patentes Baseado em Representação Vetorial Densa, Técnicas de Ordenação e Explicitação do Conhecimento / Luciano Zamperetti Wolski ; orientador, Alexandre Leopoldo Gonçalves, coorientador, José Leomar Todesco, 2023.
202 p.

Tese (doutorado) - Universidade Federal de Santa Catarina, Centro Tecnológico, Programa de Pós-Graduação em Engenharia e Gestão do Conhecimento, Florianópolis, 2023.

Inclui referências.

1. Engenharia e Gestão do Conhecimento. 2. Análise de Patente. 3. Aprendizado Profundo. 4. Classificação de Patente. 5. Grafo de Conhecimento. I. Gonçalves, Alexandre Leopoldo. II. Todesco, José Leomar . III. Universidade Federal de Santa Catarina. Programa de Pós-Graduação em Engenharia e Gestão do Conhecimento. IV. Título.

Luciano Zamperetti Wolski

Modelo de classificação de patentes baseado em representação vetorial densa, técnicas de ordenação e explicitação do conhecimento

O presente trabalho em nível de doutorado foi avaliado e aprovado no dia 1º de setembro de 2023 pela banca examinadora composta pelos seguintes membros:

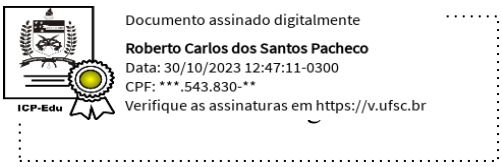
Prof. Aires José Rover, Dr.
Universidade Federal de Santa Catarina

Prof. João Artur de Souza, Dr.
Universidade Federal de Santa Catarina

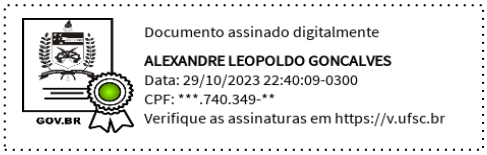
Profª. Carina Friedrich Dorneles, Dra.
Universidade Federal de Santa Catarina

Prof. Fernando Selleri Silva, Dr.
Universidade do Estado de Mato Grosso

Certificamos que esta é a versão original e final do trabalho de qualificação que foi julgado adequado para obtenção do título de doutor em Engenharia e Gestão do Conhecimento.



Coordenação do Programa de Pós-Graduação



Prof. Alexandre Leopoldo Gonçalves, Dr.
Orientador

Florianópolis, 2023.

Este trabalho é dedicado à minha mãe, Alvenir, ao meu pai, Luiz (*in memoriam*), aos meus filhos, Lucas e Maria Alice, ao enteado, João Vitor, à minha irmã, Elisane, à afilhada, Ana Clara, e à minha companheira de todas as horas, Gilvana, pelo carinho, companheirismo e compreensão durante o período de realização deste trabalho.

AGRADECIMENTOS

Em primeiro lugar, agradeço ao meu orientador, Dr. Alexandre Leopoldo Gonçalves, sua dedicação e orientação ao longo de todo o processo de pesquisa. Seu conhecimento, experiência e paciência foram fundamentais para o desenvolvimento deste trabalho.

Também expresso minha gratidão ao prof. Dr. José Leomar Todesco (Tite) por seus ensinamentos ao longo deste estudo.

Agradeço aos membros da banca Dra. Carina Friedrich Dorneles, Dr. Aires José Rover, Dr. João Artur de Souza, Dr. Fernando Selleri Silva e Dr. Douglas Alves Santos as revisões detalhadas e as valiosas sugestões que contribuíram significativamente para aprimorar este trabalho.

Minha gratidão se estende também aos colegas do Laboratório de Engenharia do Conhecimento (LEC) e aos colegas de estudo do PPGEGC que estiveram ao meu lado em todos os momentos. Nossas discussões e trocas de ideias foram enriquecedoras e motivadoras.

Acrescento o agradecimento aos colegas do PPGEGC da UFSC e, em especial, aos colegas do grupo do WhatsApp “Bolão da Mega-Sena”. Nossas conversas e interações não só tornaram esta jornada mais leve como também foram uma fonte de inspiração e motivação para mim.

Sou imensamente grato à Universidade do Estado de Mato Grosso (Unemat) e à Universidade Federal de Santa Catarina (UFSC) pelo apoio financeiro e institucional. O financiamento concedido pela Unemat foi essencial para a realização desta pesquisa, e a estrutura oferecida pela UFSC possibilitou um ambiente propício para o desenvolvimento acadêmico.

Expresso minha gratidão especial aos professores do Programa de Pós-Graduação em Engenharia, Gestão e Mídia do Conhecimento (PPGEGC) da UFSC, que me proporcionaram um ambiente estimulante e desafiador para o aprendizado e a pesquisa.

Aos professores e funcionários da Unemat que me incentivaram a fazer o doutorado, o meu muito-obrigado. Suas palavras de encorajamento e apoio foram fundamentais para minha decisão de embarcar nesta jornada acadêmica.

Por fim, minha família merece uma gratidão profunda por seu amor incondicional, apoio emocional e compreensão ao longo de todos os momentos desta trajetória. Sem o apoio e encorajamento de vocês, nada disso seria possível.

Muito obrigado a todos vocês!

RESUMO

Anualmente um grande volume de patentes é depositado nos escritórios de patentes no mundo todo, as quais precisam ser adequadamente analisadas e classificadas. Para tal, elas passam por uma avaliação detalhada conduzida por especialistas (examinadores) em um determinado domínio, de modo que possam receber um rótulo. Esse processo é custoso para os escritórios de patentes, pois o aumento no número de pedidos de patentes e a complexidade da estrutura hierárquica de categorização sobrecarrega a avaliação pelos examinadores. Ademais, a classificação precisa desses documentos é de extrema importância para a interoperabilidade entre diferentes escritórios de patentes e para a realização de tarefas confiáveis de busca, gerenciamento e recuperação de patentes durante um procedimento de pedido de patente. Portanto, é fundamental automatizar o processo de classificação, provendo meios para auxiliar os examinadores na tomada de decisão. Nesse sentido, o objetivo desta tese é propor um modelo voltado à classificação de patentes a partir de fonte de dados não estruturados na forma de texto, levando em conta aspectos de ordenação de subclasses e explicitação de conhecimento. Para cumprir esse objetivo, realizou-se uma revisão integrativa da literatura com o intuito de definir a lacuna de pesquisa e identificar os métodos e técnicas mais adequados. Após a proposição do modelo e seu desenvolvimento, este foi avaliado considerando um conjunto de dados de patentes disponibilizado pelo United States Patent and Trademark Office® (USPTO) em dois cenários, um mais geral e outro mais específico. A acurácia na avaliação do cenário geral para recomendação de subclasses ordenadas (*ranking*) ficou em torno de 80% para as três arquiteturas de redes neurais do tipo *transformers* quando consideradas as 5 (cinco) primeiras subclasses e um total de 50 documentos recuperados. No segundo cenário mais específico e com menos dados, em que o modelo foi comparado com redes neurais tradicionais na etapa de geração do *ranking*, os resultados foram mais expressivos, chegando a uma acurácia de 90%. Já quanto ao grafo de conhecimento, sua avaliação e utilização na tarefa de classificação realizada pelos examinadores, apesar de avaliações específicas não terem sido efetuadas, demonstram ser viáveis. Assim, a partir dos resultados obtidos, verifica-se que o modelo proposto permite facilitar o trabalho de examinadores na escolha de subclasses que melhor representem determinada patente.

Palavras-chave: análise de patente; classificação de patente; aprendizado profundo; *embedding;* grafo de conhecimento.

ABSTRACT

Every year, a significant volume of patents is filed with patent offices worldwide and needs to be adequately analyzed and classified. To achieve this, they undergo a detailed evaluation conducted by domain-specific experts (examiners) in order to receive a label. This process imposes a considerable burden on patent offices due to the increase in patent applications and the complexity of the hierarchical categorization structure, which overwhelms patent examination by examiners. Furthermore, the accurate classification of these documents holds paramount importance for the interoperability among different patent offices and for conducting reliable tasks of patent search, management, and retrieval during a patent application procedure. Hence, it is imperative to automate the classification process by providing means to assist examiners in decision-making. In this context, the objective of this thesis is to propose a model focused on patent classification using unstructured data sources in the form of text, taking into consideration aspects of subclass ranking and knowledge explicitation. To fulfill this objective, an integrative literature review was conducted to define the research gap and identify the most suitable methods and techniques. Following the model's proposition and development, it was evaluated using a patent dataset provided by the United States Patent and Trademark Office® (USPTO) in two scenarios: a more general one and a more specific one. The accuracy in the evaluation of the general scenario for recommending ordered subclasses (ranking) was around 80% for the three transformer-based neural network architectures when considering the top five subclasses and a total of 50 retrieved documents. In the second, more specific scenario with less data, where the model was compared with traditional neural networks in the ranking generation step, the results were more significant, achieving an accuracy of 90%. Regarding the knowledge graph, while specific evaluations were not conducted, its evaluation and utilization in the classification task performed by examiners appear to be feasible. In this sense, based on the obtained results, it is observed that the proposed model effectively facilitates the work of examiners in selecting subclasses that best represent a given patent.

Keywords: patent analysis; patent classification; deep learning; embedding; knowledge graph.

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

AAPD	Arxiv Academic Paper Dataset
AE	Auto Encoders
AI	Artificial Intelligence
ARC	Aprendizagem de Representação do Conhecimento
ANN	Artificial Neural Network
ApNN	Approximate Nearest Neighbor
BERT	Bidirectional Encoder Representations from Transformers
BIGRU	Bidirectional Gated Recurrent Unit
BI-LSTM	Bidirectional Long-Short Term Memory
BOW	Bag of Word
CAPES	Coordenação de Aperfeiçoamento de Pessoal de Nível Superior
CBOW	Continuous Bag of Word
CLEF-IP	Conference and Labs of the Evaluation Forum of Intellectual Property
CNIPA	China National Intellectual Property Administration
CNN	Convolutional Neural Network
CPC	Cooperative Patent Classification
CSV	Comma-Separated Values
CV	Computer Vision
DBOW	Distributed Bag of Words
DistilBERT	Distilled version of BERT
DL	Deep Learning
DM	Data Mining
DNN	Deep Neural Network
DS	Design Science
DSL	Deep Supervised Learning
DSR	Desing Science Research
DSRM	Design Science Research Methodology
DSSL	Deep Semi-Supervised Learning
DUL	Deep Unsupervised Learning
EC	Engenharia do Conhecimento
EGC	Engenharia e Gestão do Conhecimento

ELMo	Embeddings from Language Models
EPO	European Patent Office
EUA	Estados Unidos da América
GAN	Generative Adversarial Networks
GC	Gestão do Conhecimento
GPT	Generative Pre-trained Transformer
GRA	Gray Relational Analysis
GRU	Gated Recurrent Unit
HCI	Human-computer interaction
HFEM	Hierarchical Feature Extraction Model
ICH	Interface/Interação Homem-Computador
IE	Information Extraction
IGC	Incorporação de Grafos de Conhecimento
INPI	Instituto Nacional da Propriedade Industrial
IPC	International Patent Classification
JPO	Japan Patent Office
JSON	JavaScript Object Notation
KG	Knowledge Graph
KGE	Knowledge Graphs Embeddings
KIPO	Korean Intellectual Property Office
KNN	k-Nearest Neighbors
KR	Knowledge Representation
KRL	Knowledge Representation Learning
LDA	Latent Dirichlet allocation
LLM	Large Language Models
LSTM	Long Short-Term Memory
MARS	Multivariate Adaptive Regression Splines
ML	Machine Learning
MLP	Multilayer Perceptron
PTM	Pre-Trained Model
MyIPO	Intellectual Property Corporation of Malaysia
NER	Named Entity Recognition
NLP	Natural Language Processing

NTCIR	National Institute of Informatics Testbeds and Community for Information
OMPI	Organização Mundial da Propriedade Intelectual
OOV	Out of Vocabular
OWL	Web Ontology Language
P&D	Pesquisa e Desenvolvimento
PD&I	Pesquisa, Desenvolvimento e Inovação
PI	Propriedade Intelectual
PLN	Processamento de Linguagem Natural
PPGEGC	Programa de Pós-Graduação em Engenharia e Gestão do Conhecimento
QA	Question Answering
RBM	Restricted Boltzmann Machines
RL	Reinforcement Learning
RNN	Recurrent Neural Network
SA	Sentiment Analysis
SI	Sistemas de Informação
SQL	Structured Query Language
SVD	Singular Value Decomposition
SVM	Support Vector Machine
TF-IDF	Term Frequency-Inverse Document Frequency
TNN	Transformer Neural Network
TRIZ	Theory of Inventive Problem Solving
TRIZ-ESSL	Enhanced Semi-Supervised Learning for TRIZ
URI	Uniform Resource Identifier
USPD	United States Patented Dataset
USPTO	United States Patent and Trademark Office
WIPO	World Intellectual Property Organization
WPI	World Patent Information
XML	Extensible Markup Language

SUMÁRIO

1 INTRODUÇÃO

Os pedidos de patentes crescem anualmente e, segundo os indicadores da WIPO[1] (2022a), em 2021 foram 3,4 milhões de pedidos em todo o mundo, representando um aumento de 3,6% em relação a 2020. Pode-se dizer que esses pedidos de patentes, de uma forma bem simplificada, são grandes coleções de textos que representam uma extensa quantidade de conhecimento humano geralmente na forma não estruturada (Risch; Krestel, 2019). Todo esse conhecimento humano aplicado em um pedido de patente deve ser depositado em algum escritório de patentes espalhado por diversos países. Os escritórios de patentes fazem uma análise do pedido e, se todas as regras estipuladas forem atendidas, a patente é concedida.

Com o crescente aumento no número de pedidos de patentes depositados anualmente, as atividades exercidas pelos examinadores de patentes também cresceram. Com isso, os examinadores estão sobrecarregados, pois uma das primeiras tarefas é atribuir manualmente códigos de classificação a essas patentes com base em seu conteúdo técnico (Sofean, 2021).

O desenvolvimento contínuo da tecnologia aumenta o número de novas invenções e, consequentemente, o número de patentes. Cada nova tecnologia é classificada em uma das subclasses de patentes existentes. Caso não se enquadre em nenhuma subclasse existente, uma nova subclasse de patente é criada. Esses processos apoiam o desenvolvimento de métodos automatizados de classificação de patentes, que, nos últimos anos, têm sido conduzidos no campo da classificação automática de patentes (Yücesoy Kahraman; Dereli; Durmuşoğlu, 2023).

Dessa forma, assegura-se que as patentes e os pedidos de patentes com características semelhantes, que tratem de temas similares ou de áreas tecnológicas específicas, sejam agrupados sob os mesmos códigos. A correta classificação de documentos de patentes é de extrema importância para a interoperabilidade entre diferentes escritórios de patentes e para a realização de buscas de maneira confiável, durante um procedimento de pedido de patente (Gomez; Moens, 2014).

O sistema de classificação utilizado pela maioria dos escritórios de patentes é a International Patent Classification (IPC), estabelecida pelo Acordo de Estrasburgo no ano de 1971. A IPC consiste em um sistema hierárquico que divide as tecnologias em oito seções que vão de A a H e cerca de 75 mil subdivisões, cada uma representada por um símbolo que consiste em caracteres do alfabeto latino e algarismos arábicos (WIPO, 2022b).

[1] WIPO é o acrônimo de World Intellectual Property Organization ou Organização Mundial da Propriedade Intelectual.

A IPC classifica sistematicamente as patentes em subclasses auxiliando na recuperação de documentos semelhantes, que pode ser realizada através da busca por patentes na mesma categoria. Entretanto, classificar patentes manualmente se torna dispendioso aos examinadores e requer conhecimento específico do domínio devido à complexidade do IPC (Risch; Krestel, 2019).

Nesse sentido, a Inteligência Artificial, por meio de técnicas de aprendizado de máquina, e, mais recentemente, o aprendizado profundo estão entre as abordagens mais utilizadas para solucionar as mais variadas tarefas na análise de patentes, em especial a classificação de patentes. O uso de sistemas automatizados de classificação de patentes pode ser muito eficaz, uma vez que esses sistemas suportam uma variedade de atividades de inovação tecnológica, incluindo examinar, detectar e reduzir a possibilidade de violação de patentes (Yun; Geum, 2020).

No entanto, quanto maior o volume de patentes, mais complexo se torna o processo de classificação. Segundo Gomez e Moens (2014), quando se tem muitas patentes para gerenciar, a estrutura de classificação deve ser muito bem organizada e detalhada para facilitar a classificação, a navegação e a busca precisa. Além disso, como as patentes refletem de alguma forma o conhecimento tecnológico do mundo e como esse conhecimento muda ao longo do tempo, a estrutura de classificação também deve ser flexível o suficiente para capturar tais mudanças.

Há vários desafios na classificação automática de patentes. Um dos principais é a complexidade da linguagem utilizada nos documentos de patentes, que muitas vezes contêm terminologias técnicas e jurídicas altamente especializadas. Além disso, a criação de novas tecnologias e a constante evolução das já existentes dificultam manter as classificações atualizadas. Outro desafio é o grande volume de patentes que precisam ser classificadas, tornando o processo demorado e caro. Apesar do desenvolvimento de técnicas de processamento de linguagem natural e aprendizado profundo, a classificação automática de patentes ainda não é tão precisa quanto a classificação manual realizada por especialistas no assunto (Yoo *et al*., 2023).

A classificação de patentes, como tarefa, possui vários desafios que impactam diretamente na tomada de decisão por parte dos examinadores. Com isso, a classificação de patentes, caso forneça os elementos adequados, pode auxiliar no aumento da qualidade da tomada de decisão efetuada pelos examinadores nos escritórios de patentes (Jafery *et al.*, 2019).

1.1 IDENTIFICAÇÃO DO PROBLEMA

A capacidade de inovação caracteriza um ativo estratégico fundamental para as empresas fornecerem e manterem sua vantagem competitiva. Representa a capacidade das organizações de estimular e de criar conhecimentos aplicando tecnologias apropriadas na gestão dos seus processos (Ponta *et al.*, 2020).

Ademais, as inovações desempenham um papel importante no desenvolvimento de qualquer país tanto em termos de tecnologia quanto de economia, implicando na criação e na transferência de conhecimento para um novo produto ou processo (Girthana; Swamynathan, 2020).

No contexto da inovação, a análise de patentes desempenha um papel fundamental nas organizações, pois lida com i) uma grande quantidade de informações sobre o progresso tecnológico e as tendências do mercado que podem conduzir a decisões de investimento (Ploskas *et al.*, 2019) e com ii) uma variedade de informações sobre inovação tecnológica amplamente utilizadas em estudos de inovação e gerenciamento de tecnologia (Noh; Lee, 2020).

Um documento de patente é um indicador das atividades de inovação na prática. As patentes são usadas como insumos para acompanhar as tendências tecnológicas de um setor. Entende-se também que a análise de patentes é uma ferramenta importante para avaliar a capacidade de investimento, gerenciamento de concorrência e planejamento de gestão de negócios em setores intensamente competitivos (Altuntas, 2023).

De modo geral, a análise de patentes permite que equipes de PD&I fomentem, por exemplo, o conhecimento das últimas tendências tecnológicas com o intuito de acelerar a produção (Girthana; Swamynathan, 2020).

Todavia, o rápido crescimento no volume de patentes torna-se um desafio, exigindo o desenvolvimento de ferramentas adequadas e essenciais para as seguintes ações: i) analisar e prever tendências tecnológicas futuras (Choi; Hong, 2020; Evangelista *et al.*, 2020); ii) realizar planejamento estratégico de tecnologia (Girthana; Swamynathan, 2020; Yu; Zhang, 2019); iii) detectar violação de patentes (Girthana; Swamynathan, 2020; Sorce *et al.*, 2018); iv) determinar a qualidade de patentes (Lo; Cho; Wang, 2020; Wu, 2019); e v) identificar patentes mais promissoras (Geum; Kim, 2020; Noh; Lee, 2020).

As patentes são avaliadas de acordo com a sua importância ou potencial. Sendo assim, a classificação de patentes é uma tarefa desafiadora e relevante que, se executada adequadamente, pode auxiliar os tomadores de decisão a investir em novas soluções industriais

inovadoras. A classificação de patentes pode aprimorar a qualidade das decisões estratégicas baseadas em patentes, bem como elevar o valor da inovação, ao mesmo tempo que reduz as violações de patentes (Jafery *et al.*, 2019).

A tarefa de classificar patentes é uma forma de gerir o conhecimento que possibilita que os documentos sejam atribuídos a categorias predefinidas. Os textos dos documentos de patentes são extensos, redigidos em linguagem muito técnica, compostos por vocabulário vasto e palavras ruidosas, que reduzem o desempenho da análise em termos de precisão. Uma das tarefas do examinador de patentes é atribuir manualmente códigos da IPC às patentes de acordo com o conteúdo técnico (Sofean, 2021).

A classificação de patentes é tipicamente uma tarefa realizada quase de forma exclusiva por examinadores de patentes, sendo considerada a tarefa mais fundamental da análise de patentes. Além disso, a classificação de patentes enfrenta vários desafios. Em primeiro lugar, a taxonomia do IPC é uma estrutura hierárquica intrincada. Cada patente deve receber um ou mais rótulos de nível de subgrupo. Em segundo lugar, a distribuição de patentes entre as categorias é altamente desequilibrada, com cerca de 80% de todos os documentos classificados em cerca de 20% das categorias. Além disso, os documentos de patentes costumam ser longos e cheios de terminologias técnicas e jurídicas, o que aumenta o desafio de uma análise eficiente, mesmo para especialistas no domínio (Li *et al.*, 2018).

Segundo Ernst (2003), as informações sobre patentes devem se tornar o elemento central de um sistema de gestão do conhecimento numa empresa. Dessa forma, a recuperação e a avaliação de dados de gestão do conhecimento devem ser institucionalizadas dentro da organização, a fim de garantir o uso contínuo e sistemático de informações de patentes nos processos de tomada de decisão de uma empresa.

A gestão de patentes forma ativos intangíveis para a tecnologia da empresa ao mesmo tempo que apoia a gestão da inovação. Dessa maneira, uma gestão bem-sucedida deve buscar simultaneamente duas perspectivas, ou seja, uma visão interna das patentes da própria empresa e uma visão externa das patentes de terceiros. Com isso, a gestão de patentes tornou-se uma função central em empresas orientadas para a tecnologia, compreendendo vários componentes que requerem decisões gerenciais sobre como ou em que grau elas devem ser implementadas (Moehrle; Walter; Wustmans, 2017).

Segundo Gassmann, Bader e Thompson (2021), um componente essencial da gestão de patentes é a condução de procedimentos no processo de pedido de patente perante os escritórios de patentes (processo de patente). Recentemente, alguns autores convergiram para a visão de gestão de patentes composta por processos centrais (ou seja, geração de patentes,

gestão de portfólio, inteligência, exploração e aplicação) e dimensões de suporte (estratégia e organização para patenteamento) (Agostini; Nosella; Holgersson, 2023; Agostini; Nosella; Teshome, 2019; Moehrle; Walter; Wustmans, 2017).

De acordo com Somaya (2016), a gestão de patentes vem ganhando cada vez mais importância nas organizações. O gerenciamento adequado de patentes permite que as empresas se apropriem do valor das atividades de Pesquisa, Desenvolvimento e Inovação (PD&I) e obtenham uma vantagem competitiva sustentável. Dessa forma, um processo de gestão de patentes eficaz pode levar a alguns benefícios, tais como alavancar o valor inerente da patente por meio da definição de uma estratégia de exploração apropriada, economizando recursos financeiros, e equilibrar o custo associado ao depósito e manutenção de patentes com capacidade de gerar lucros para a empresa (Soranzo; Nosella; Filippini, 2016).

Segundo Agostini, Nosella e Holgersson (2023), a gestão de patentes envolve vários processos fundamentais como geração, gerenciamento de portfólio, exploração e inteligência de patentes. É um processo complexo que requer uma perspectiva estratégica e uma boa organização para desenvolver processos sofisticados. O uso da estratégia de patentes, apoiada por uma organização e cultura de patentes bem desenvolvidas, influenciará positivamente os processos de gerenciamento do portfólio de patentes de uma empresa.

Atualmente, não há um sistema automático de classificação comum usado pelos escritórios de patentes e nenhuma abordagem abrangente para classificar automaticamente todo o conjunto de patentes disponíveis no escritório de patentes. O grande número de níveis da hierarquia do IPC causa problema de escalabilidade e torna a tarefa de classificação computacionalmente cara (Roudsari *et al*., 2020).

De acordo com Yoo *et al*. (2023), atualmente utilizam-se técnicas de processamento de linguagem natural e aprendizado profundo para classificação automática de patentes. O objetivo é atribuir automaticamente um ou mais códigos IPC a um pedido de patente. Isso é conduzido por meio de algoritmos, responsáveis por analisar o texto do documento de patente e identificar palavras-chave e frases que indicam a área técnica da invenção. Essas informações são então usadas para determinar os códigos IPCs apropriados.

Apesar disso, a busca pela classificação automática de patentes ainda é um desafio devido ao rápido crescimento no número de pedidos de patentes e ao surgimento de novos tipos de inovação (Cassidy, 2020; Jafery *et al.*, 2019; Risch; Krestel, 2019; Roudsari *et al.*, 2020; Yoo *et al*., 2023; Yun; Geum, 2020). Para Risch e Krestel (2019), a grande coleção de textos não estruturados de patentes torna a extração de informações um desafio e a classificação e a recuperação automática de documentos tarefas difíceis de serem realizadas.

Além disso, a explicitação do conhecimento latente contido nas patentes requer a identificação e a representação do conhecimento técnico de forma estruturada. Explicitar o conhecimento sobre patentes significa tornar o conhecimento contido no documento de patente acessível, compreensível e utilizável. Isso envolve a organização e a estruturação das informações armazenadas nas patentes, de modo a facilitar a busca, a análise e a aplicação desse conhecimento (Becattini *et al*., 2015).

A explicitação do conhecimento é um componente essencial da Gestão do Conhecimento (GC), pois auxilia o compartilhamento do conhecimento, a colaboração e o aprendizado organizacional (Baskerville; Dulipovici, 2006). A explicitação, nesse contexto, tem o papel de auxiliar na classificação automática, permitindo uma melhor representação do conteúdo técnico das patentes.

Diante do exposto, percebe-se a necessidade de melhoria e agilidade no processo para classificar patentes. Sendo assim, é essencial observar questões relacionadas à classificação de patentes que promovem desafios nesse contexto, tais como:

- Volume de dados: aproximadamente 3 milhões de pedidos de patentes são depositados todos os anos (WIPO, 2022a). Isso gera um problema de escalabilidade (Meireles; Ferraro; Geva, 2016; Roudsari *et al*., 2020), bem como um alto custo computacional para classificar patentes (Bai; Shim; Park, 2020) devido ao grande volume de dados;
- Gestão de patentes: visto que a recuperação e a avaliação de dados de patentes são fundamentais com o intuito de garantir o uso contínuo e sistemático de informações de patentes nos processos de tomada de decisão de uma empresa (Agostini; Nosella; Holgersson, 2023; Agostini; Nosella; Teshome, 2019; Ernst, 2003; Gassmann; Bader; Thompson, 2021; Moehrle; Walter; Wustmans, 2017; Somaya, 2016; Soranzo, Nosella, Filippini, 2016);
- Classificação de patentes: devido à complexidade da taxonomia da IPC/CPC (Cooperative Patent Classification*)*, a distribuição desequilibrada das patentes entre as categorias (Li *et al*., 2018) e a reclassificação de patentes que ocorre anualmente;
- Classificação automática de patentes: uma vez que os escritórios de patentes não possuem um sistema de classificação automática comum entre eles (Roudsari *et al*., 2020). Ademais, a grande coleção de textos não estruturados promove desafios

na extração de informações, tornando a classificação e a recuperação automática de documentos tarefas complexas (Risch; Krestel, 2019).

Diante do exposto, é inerente considerar questões como o volume de dados e a gestão de patentes quando se trata do cenário de classificação de patentes, principalmente devido à importância e ao impacto dessa atividade nas organizações. No entanto, ressalta-se que os desafios centrais que motivaram esta tese, estando no cerne da pesquisa, são os relacionados especificamente à tarefa de classificação de patentes, o que inclui problemas como a complexa taxonomia IPC, a distribuição irregular de patentes entre categorias, a linguagem técnica dos documentos, a ausência de sistemas unificados de classificação automática e a constante evolução tecnológica.

1.2 PERGUNTA DE PESQUISA

Diante do contexto apresentado, surge a seguinte pergunta de pesquisa: como auxiliar na análise de patentes, mais especificamente na tarefa de classificação, por meio de elementos que caracterizem a relevância de determinada categoria e que explicitem o conhecimento latente presente em grandes bases de dados de patentes?

1.3 OBJETIVOS

1.3.1 Objetivo geral

Propor um modelo voltado à classificação de patentes a partir de grandes volumes de dados não estruturados, levando em conta aspectos de ordenação de subclasses e explicitação de conhecimento.

1.3.2 Objetivos específicos

Os objetivos específicos definidos para a presente tese são os que se seguem:

- identificar e desenvolver estratégias de recomendação de subclasses considerando aspectos de ordenação e relevância dessas subclasses;
- analisar e desenvolver métodos e técnicas que permitam explicitar o conhecimento latente em bases de patentes a partir da recomendação de subclasses; e

- elaborar diretrizes e cenários de estudo baseados em grandes volumes de dados não estruturados para aplicação do modelo em contextos reais de recomendação de subclasses e explicitação do conhecimento.

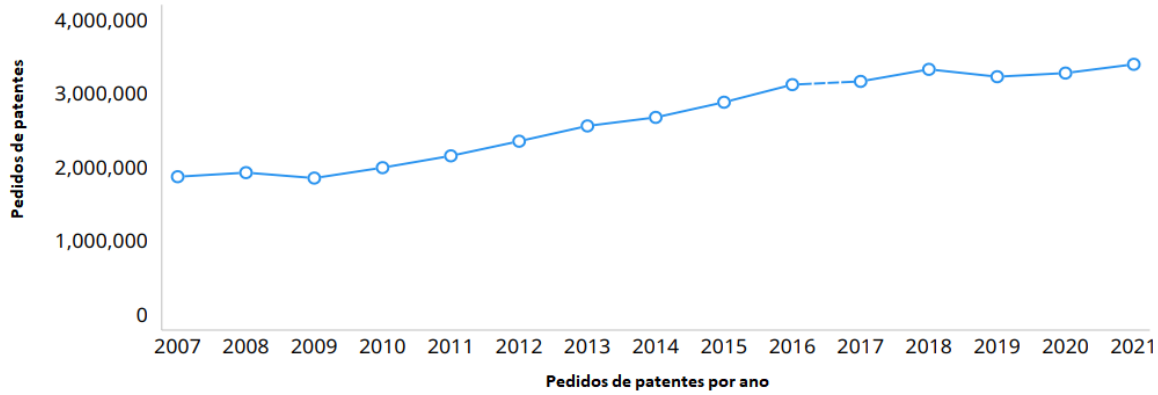
1.4 JUSTIFICATIVA E RELEVÂNCIA DO TEMA

As patentes possuem importante papel em uma economia baseada no conhecimento, uma vez que as empresas usam tal recurso para proteger a inovação e o domínio do mercado de novos produtos por um período de tempo (Trappey *et al.*, 2012).

Com o crescente aumento no número de pedidos de patentes depositados anualmente, aumenta também a necessidade de sistemas eficazes e eficientes que satisfaçam o gerenciamento dessa grande quantidade de dados, de modo que sejam ferramentas importantes para a análise de patentes (Shalaby; Zadrozny, 2019).

Segundo levantamento da WIPO (WIPO, 2022a), em todo o mundo foram requeridos 3,4 milhões de pedidos de patentes em 2021, o que representa um aumento de 3,6% em relação a 2020. Em 2019, houve uma redução de 3% em relação ao ano de 2018, constituindo a primeira queda desde a crise financeira de 2009. Mesmo com a redução no número de pedido de patentes em 2019, o volume de dados voltou a crescer nos anos de 2020 e 2021. A Figura 1 apresenta o valor acumulado nos últimos 15 anos (WIPO, 2022a).

Figura 1 – Pedidos de patentes em todo o mundo de 2007 a 2021



Fonte: adaptado de WIPO (2022a)

Segundo List (2018), o aumento no número de patentes publicadas tem impulsionado o desenvolvimento de plataformas de análise para revisão desses documentos em grande escala.

Tais plataformas oferecem ferramentas de visualização, categorização tecnológica e indicadores de valor, auxiliando os tomadores de decisão.

Para Moehrle *et al.* (2010), o processo de análise de patentes possui três etapas: 1) pré-processamento, 2) processamento e 3) pós-processamento. Segundo Aristodemou e Tietze (2018), na etapa de pré-processamento os dados são coletados, e após a extração das informações ocorre a limpeza e a preparação com o objetivo de disponibilizar esses dados com qualidade, exatidão e completude. Na etapa de processamento, a análise dos dados extraídos na etapa de pré-processamento é realizada por meio de diferentes métodos para classificar, agrupar e identificar percepções significativas a partir das informações. Por fim, na etapa de pós-processamento, os resultados e as informações da etapa de processamento são visualizados e avaliados para apoiar a tomada de decisão estratégica.

As tarefas típicas de análise de patentes incluem: i) exploração de tecnologia a fim de capturar tecnologias novas e modernas em um domínio específico e, subsequentemente, usá-las para criar serviços inovadores; ii) análise do cenário de tecnologia para avaliar a densidade de depósitos de patentes de tecnologia específica e, subsequentemente, atividades de PD&I; (iii) análise competitiva e *benchmarking* para identificar pontos fortes e diferenças do portfólio de patentes da empresa em comparação com outros participantes importantes que trabalham com tecnologias relacionadas; iv) classificação e pontuação de patentes para quantificar a força das reivindicações de uma patente existente ou nova; e v) pesquisa da técnica anterior com o objetivo de recuperar documentos de patentes e outras publicações científicas relevantes para um novo pedido de patente. Portanto, as atividades relacionadas à análise de patentes requerem um alto nível de conhecimento de domínio que, mesmo se disponível, deve ser integrado com análises altamente sofisticadas e inteligentes que forneçam assistência cognitiva e interativa aos usuários (Shalaby; Zadrozny, 2019).

Com isso, após o depósito do pedido de patente em um escritório de patentes, um examinador qualificado avalia a patenteabilidade da invenção descrita. O examinador verifica se a invenção reivindicada é nova e inventiva em relação ao estado da técnica preexistente. A esse respeito, o examinador cita principalmente patentes publicadas mais antigas ou publicações de patentes em vez de livros ou artigos de conferências como técnica anterior. As principais razões para citar outras patentes e pedidos de patentes são o grande volume de documento de patentes, a estrutura padronizada dos dados em classes de patentes e a data de publicação inquestionável de cada documento de patente (Krestel *et al.*, 2021).

Em pesquisas de patentes para avaliação da técnica anterior, um especialista/examinador em tecnologia precisa determinar a palavra-chave para a tecnologia-

alvo, pesquisar patentes e recuperar o conjunto de documentos de patentes do banco de dados de suporte. Nesse processo, o resultado inicial da recuperação ainda é uma mistura de documentos de patentes que são relevantes e irrelevantes para a tecnologia-alvo a ser analisada. Portanto, os examinadores devem classificar a saída para obter apenas os documentos de patentes relacionados (Kim *et al.*, 2020).

Portanto, uma das tarefas mais frequentes do processamento de patentes em um escritório de patentes é a classificação de patentes. Essa tarefa tem o objetivo de classificar os textos de patentes em várias categorias, confirmando ou acrescentando outras categorias definidas previamente no pedido de patente (Korde; Mahender, 2012).

Cada patente possui elementos textuais que indicam sua área técnica e seu campo de invenção. O título geralmente sintetiza a ideia central, enquanto o resumo apresenta as informações essenciais sobre a invenção. O relatório descritivo detalha o funcionamento, as elucidações e as reivindicações. A análise desses componentes possibilita identificar os códigos mais adequados da IPC para categorizar determinada patente, de maneira hierárquica, em seções, classes, subclasses, grupos e subgrupos. Assim, os textos das patentes contêm termos e conceitos que, se adequadamente extraídos e processados, podem apoiar a tarefa de classificação conduzida pelos examinadores segundo a taxonomia IPC (Risch; Krestel, 2019). No Anexo A, tem-se um exemplo da patente US08472379 publicada.

A IPC provê uma estrutura lógica para organizar o conhecimento tecnológico patenteado em oito amplas áreas, de A a H, facilitando a busca e a recuperação de patentes, conforme o Quadro 1 e o Apêndice A.

Quadro 1 – Seções da taxonomia IPC

Seção	Descrição
A	Necessidades Humanas
B	Operações de Processamento; Transporte
C	Química; Metalurgia
D	Têxteis; Papel
E	Construções Fixas
F	Engenharia Mecânica; Iluminação; Aquecimento; Armas; Explosão
G	Física
H	Eletricidade

Fonte: elaborado pelo autor (2023)

Os escritórios de patentes classificam as patentes de acordo com a International Patent Classification (IPC), sistema hierárquico criado com o objetivo de classificar as invenções e seus documentos em áreas técnicas que abrangem todas as áreas da tecnologia (Meguro; Osabe, 2019). A IPC é uma estrutura complexa e hierárquica, compreendendo 8 seções, 128 classes,

648 subclasses, cerca de 7.200 grupos principais e aproximadamente 72.000 subgrupos (Yun; Geum, 2020), organizada conforme o Quadro 2:

Quadro 2 – Exemplo de classificação de patente para a IPC

A41D 25/02		
Seção	A	Necessidades Humanas
Classe	A41	Vestuário
Subclasse	A41D	Agasalhos; Vestuário de proteção; Acessórios
Grupo principal	A41D 25/00	Gravatas
Subgrupo	A41D 25/02	Com nó ou laço já feito, com ou sem a volta do pescoço

Fonte: elaborado pelo autor (2023)

A classificação de patentes é uma tarefa fundamental na análise de patentes, sendo tradicionalmente realizada por especialistas e examinadores de patentes. No entanto, enfrenta diversos desafios. A taxonomia da IPC é uma estrutura hierárquica complexa, em que cada patente precisa ser categorizada em um ou mais rótulos de nível de subgrupo. Além disso, a distribuição de patentes entre as categorias é altamente desequilibrada, com a maioria dos documentos classificados em apenas algumas categorias. Os documentos de patentes são extensos e contêm terminologias técnicas e jurídicas, o que dificulta uma análise eficiente mesmo para especialistas no domínio. Atualmente, não há um sistema de classificação automática padronizado usado pelos escritórios de patentes, tornando a classificação de todo o conjunto de patentes disponíveis computacionalmente dispendiosa devido ao grande número de níveis da hierarquia da IPC (Fall; Benzineb, 2002; Li *et al*., 2018; Roudsari *et al*., 2020).

Além disso, devido ao rápido crescimento no número de pedidos de patentes, a reclassificação anual de patentes, o surgimento de novos tipos de inovação e a classificação automática de patentes (Cassidy, 2020; Jafery *et al*., 2019; Risch; Krestel, 2019; Roudsari *et al.*, 2020; Yun; Geum, 2020) se tornam um desafio para os examinadores de patentes.

Com o objetivo de identificar uma lacuna de conhecimento a ser preenchida nesse tema, foi realizada uma revisão integrativa da literatura. A partir da revisão não foram encontrados estudos com a mesma combinação de técnicas e recursos proposta neste trabalho. Para preencher a lacuna identificada, serão combinadas técnicas de processamento de linguagem natural, representação do conhecimento e aprendizado profundo. Pretende-se apresentar um modelo capaz de recomendar subclasses de patentes de forma ordenada e também explicitar conhecimento latente em bases de patentes considerando tais recomendações.

Para isso, propõe-se a criação de uma base de conhecimento para buscas semânticas que possibilite comparações por similaridade vetorial, assim como a utilização de grafos de

conhecimento relacionando subclasses e conceitos extraídos das patentes. O modelo terá uma natureza dinâmica, incorporando continuamente novos documentos de patentes à base de conhecimento. Será implementado um protótipo para demonstração e aplicação, com foco no fluxo de incorporação de patentes. Nesse sentido, ainda que a acurácia nas recomendações seja relevante para o modelo, o foco reside na oferta de ferramental que auxilie examinadores na tarefa de classificação de patentes.

Dessa forma, o trabalho se justifica por preencher uma lacuna da literatura, combinando técnicas ainda não exploradas conjuntamente e incorporando recursos de ordenação, explicitação e representação do conhecimento ao contexto.

1.5 ORIGINALIDADE DA PESQUISA

Nesta pesquisa é proposto um modelo que estabelece uma combinação de técnicas de processamento de linguagem natural, representação do conhecimento e aprendizado profundo para auxiliar no processo de classificação de patentes. Mais especificamente, o trabalho pretende contribuir com o processo de tomada de decisão desempenhado por examinadores, de tal maneira que a tarefa de classificação no contexto de análise de patentes seja facilitada.

Com o objetivo de identificar o problema e propor uma solução, é importante realizar uma busca na literatura para apontar uma lacuna de conhecimento que relacione métodos e técnicas de Engenharia do Conhecimento (EC) para classificar patentes de forma mais adequada com o intuito de auxiliar na tomada de decisão. As técnicas de EC envolvem a aquisição, a representação, o raciocínio, o aprendizado, a explicação e a validação para a construção de sistemas baseados em conhecimento (Nazário; Dantas; Todesco, 2014). As principais técnicas computacionais e de EC contempladas neste trabalho são:

- Processamento de linguagem natural (NLP): para extração de informações relevantes dos textos de patentes;
- Aprendizado profundo: técnicas como redes neurais para apoiar a sugestão de subclasses para patentes;
- *Embeddings*: representação vetorial do conteúdo das patentes para comparação de similaridade;
- Grafo de conhecimento: para modelar e explicitar relações entre conceitos e subclasses extraídos das patentes.

Dessa forma, técnicas relacionadas à representação, à modelagem e ao processamento de conhecimento a partir de dados textuais são contempladas neste trabalho, visando apoiar examinadores na tarefa de classificação de patentes.

Para certificar a originalidade e a relevância do tema, foi realizada uma revisão integrativa da literatura a fim de se obter uma melhor compreensão sobre o tema proposto. Dessa forma, desenvolveu-se uma busca na literatura nas bases de dados Science Direct®, Scopus®, Web of Science® (WoS) e IEEE®. Essas bases de dados foram escolhidas devido à sua credibilidade e relevância no âmbito acadêmico. No capítulo 3 serão apresentados os detalhes metodológicos utilizados nessa revisão. A partir da revisão integrativa da literatura, foi possível constatar que:

- dos estudos encontrados, todos afirmam ser eficazes para a classificação de patentes utilizando-se de demonstração e de estudos de caso. Percebe-se que a melhoria da eficiência e precisão na análise de patentes está ligada ao uso de técnicas de Inteligência Artificial (do inglês *Artificial Intelligence* - AI) para agilizar o processo de classificação de patentes. Entre as subáreas/subcampos mais utilizadas de AI nesse contexto estão o aprendizado de máquina e o aprendizado profundo;
- nos últimos anos, técnicas de aprendizado de máquina combinadas com aprendizado profundo estão entre as mais utilizadas. No que tange à proposta desta tese, pretende-se apresentar um modelo que recomende uma relação de subclasses ordenadas (*ranking*) de patentes, bem como meios de auxiliar os examinadores por meio da explicitação do conhecimento. Na revisão integrativa da literatura não foram encontrados trabalhos que demonstrassem tal possibilidade;
- percebe-se o esforço nas pesquisas para classificar patentes de forma automática e com maior precisão, combinando várias técnicas e algoritmos que auxiliem na busca e na recuperação de documentos de patentes. Os métodos e as técnicas de processamento de linguagem natural se mostram eficientes na tarefa de classificar patentes, mesmo que não se tenha um procedimento único para tal;
- através da revisão integrativa da literatura, percebe-se que os métodos e as técnicas de Engenharia do Conhecimento são fundamentais para a tomada de decisão, podendo auxiliar examinadores na classificação adequada de patentes. Ou seja, ao chegar uma nova patente, o ferramental ofertado ao examinador deve prover suporte adequado para a realização da tarefa de classificá-la.

Sendo assim, a criação de um modelo que auxilie na classificação de patentes com o intuito de tornar a realização da tarefa mais assertiva se faz necessária, de modo a minimizar custos e recursos. Nesse contexto, foi identificada uma lacuna na literatura que será tratada de forma inédita, ou seja, através da proposição de um modelo voltado à recomendação de subclasses de patentes a partir de fontes de dados não estruturados na forma de texto, levando em conta aspectos de ordenação de subclasses (*ranking*) e explicitação de conhecimento.

1.5.1 Contribuições

Esta tese tem como principal contribuição a proposição de um modelo para auxiliar no processo de classificação de patentes. O modelo constitui-se de uma combinação de métodos e de técnicas da Engenharia do Conhecimento com o intuito de suportar a análise de patentes, mais especificamente a tarefa de classificação.

Além da contribuição principal, outras contribuições da pesquisa podem ser elencadas:

- criação de uma base de conhecimento voltada à comparação por similaridade como suporte à recomendação de subclasses;
- criação de uma estrutura (grafo de conhecimento) relacionando subclasses e conceitos (tópicos) com o intuito de explicitar o conhecimento durante a classificação de patentes;
- dinamicidade do modelo proposto, incorporando os novos documentos de patentes à base de conhecimento e ao grafo de conhecimento; e
- implementação de um protótipo para demonstrar a viabilidade do modelo proposto bem como a aplicação em um cenário de estudo.

1.6 ESCOPO DO TRABALHO

Esta tese propõe um modelo voltado à tarefa de classificação de patentes por meio de um conjunto de elementos que auxilie na análise de patentes, de modo que esta possa ser realizada de maneira adequada. Sendo assim, para se classificar adequadamente uma patente, utilizam-se métodos e técnicas da Engenharia do Conhecimento.

O modelo proposto emprega uma combinação de técnicas de processamento de linguagem natural, aprendizado profundo e explicitação de conhecimento para classificar ou, mais especificamente, recomendar subclasses de maneira ordenada, ou seja, através de *ranking*,

e a partir disso disponibilizar meios de explicitar o conhecimento envolto nas sugestões. Nesse sentido, o foco não está em criar um novo classificador, tampouco buscar níveis elevados de acurácia na classificação, apesar de esta tese analisar os resultados do modelo visando gerar subsídios para a sua avaliação. O foco da tese reside principalmente no fluxo desde a captura dos dados (patentes), passando pela explicitação de conhecimento, até a incorporação do resultado da classificação, de tal maneira que a tarefa de classificação, no contexto da análise de patentes, auxilie os examinadores na tomada de decisão.

O conjunto de dados de patentes será da United States Patent and Trademark Office (USPTO) disponível na *web* no idioma inglês, mais precisamente o conjunto de dados de referência para classificação de patentes, denominado de USPTO-2M (Li *et al*., 2018). O USPTO-2M possui mais de dois milhões de registros de patentes de utilidade nos Estados Unidos. O modelo proposto nesta tese utiliza esse conjunto de dados, os quais serão divididos para as etapas de treinamento e teste.

Para a avaliação do modelo proposto, faz-se uso de métrica de avaliação fundamentada na tarefa de classificação de textos, a qual tem o intuito exclusivo de avaliar o desempenho da etapa de aprendizado durante a instanciação do modelo proposto, mais especificamente a recomendação de subclasses. O modelo desta tese não objetiva informar com precisão absoluta que determinada patente deva ser classificada em A, B ou C. Objetiva sim sugerir uma relação ordenada (*ranking*) de subclasse de patentes, assim como prover meios de explicitar o conhecimento do domínio, visando auxiliar examinadores na tomada de decisão, ou seja, durante a atribuição de subclasses a determinada patente.

1.7 ADERÊNCIA AO PPGEGC

O objetivo desta seção é apresentar o contexto da tese e destacar a aderência do trabalho proposto ao Programa de Pós-Graduação em Engenharia, Gestão e Mídia do Conhecimento (PPGEGC). Para tal, a seção foi dividida em duas subseções, sendo que uma apresenta a identidade da tese e a outra as referências factuais relacionadas à pesquisa.

1.7.1 Identidade

A aderência da proposta de tese ao PPGEGC pode ser contextualizada com o objetivo do Programa, que é “formar pesquisadores comprometidos com o ensino, a pesquisa e o

desenvolvimento voltados à codificação, gestão e disseminação do conhecimento nas organizações e sociedade" (EGC, 2019).

Perante os objetivos do PPGEGC, esta tese é aderente ao Programa por desenvolver um modelo de conhecimento sobre conteúdo técnico/científico, formalizado e codificado na forma de um sistema (nível protótipo) que combina ferramentas para a resolução de problemas, mais especificamente no auxílio à classificação de patentes.

A área de concentração da tese é a Engenharia do Conhecimento (EC), com linha de pesquisa em Teoria e Prática em Engenharia do Conhecimento. Essa linha de pesquisa aborda metodologias e tecnologias da EC e da Inteligência Computacional e suas relações com a Gestão e com a Mídia do Conhecimento (EGC, 2019)[2].

A definição de conhecimento para o PPGEGC ocorre a partir da interdisciplinaridade, em que cada uma das áreas de conhecimento do Programa possui sua visão de mundo, com diferentes escolas científicas. Do ponto de vista epistemológico, na EC a principal referência é a visão cognitivista, na qual o conhecimento é identificado como uma entidade que pode ser armazenada em computadores, bases de dados, arquivos, manuais ou rotinas para posterior compartilhamento em uma organização (Pacheco, 2016).

Em relação à interdisciplinaridade, foi utilizado o entendimento do documento de área interdisciplinar da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES):

> Entende-se por Interdisciplinaridade a convergência de duas ou mais áreas do conhecimento, não pertencentes à mesma classe, que contribua para o avanço das fronteiras da ciência e tecnologia, transfira métodos de uma área para outra, gerando novos conhecimentos ou disciplinas e faça surgir um novo profissional, com um perfil distinto dos existentes, com formação básica sólida e integradora, capaz de compreender e solucionar os problemas cada vez mais complexos das sociedades modernas (CAPES, 2019, p. 9).

Nesse ambiente interdisciplinar, as pesquisas desenvolvidas na área de EC tratam de desenvolver técnicas, metodologias e ferramentas que apoiem as atividades da Gestão do Conhecimento (GC) (Pacheco, 2014). Os processos da GC abordados nesta tese são criação, compartilhamento e armazenamento do conhecimento. Já a Gestão do Conhecimento "é a gestão das atividades e processos organizacionais que promovem o conhecimento organizacional para o aumento da competitividade, por meio do melhor uso e da criação de fontes de conhecimento individuais e coletivas" (CEN, 2004).

[2] Disponível em: https://ppgegc.paginas.ufsc.br/areas-de-concentracao. Acesso em: 4 nov. 2021.

Sendo assim, esta tese procura prover meios para auxiliar uma tarefa com impacto na gestão do conhecimento organizacional, que é a classificação de patentes, pois envolve pessoas que estabelecerão em qual categoria uma determinada patente será inserida. Isso depende muito do conhecimento do especialista de patentes, nesse contexto tratado como examinador. Portanto, a GC se utiliza de ferramental provido pela EC para suportar examinadores de patentes, criando um fluxo voltado à tomada de decisão.

1.7.2 Referências factuais

Na área de análise de patentes, mais especificamente na classificação de patentes, não foram encontrados registros no Banco de Teses e Dissertações do PPGEGC sobre o tema. Sendo assim, os temas selecionados mais pertinentes a esta tese são análise de redes, aprendizado de máquina, classificação de documentos, classificação de texto, descoberta de conhecimento, mineração de dados e mineração de texto. Entre os temas selecionados, identificaram-se 14 trabalhos, a saber:

1) Análise de redes
   - Silva, Lucyene Lopes da. Framework conceitual Dandelion de análise de redes sociais e tecnologias da informação e comunicação para organizações em rede. Tese - Programa de Pós-graduação em Engenharia, Gestão e Mídia do Conhecimento, Universidade Federal de Santa Catarina, Florianópolis, 2020.
   - Formanski, José Gilberto. A estrutura da rede social organizacional e sua influência no fluxo de conhecimento inovador. Tese - Programa de Pós-graduação em Engenharia, Gestão e Mídia do Conhecimento, Universidade Federal de Santa Catarina, Florianópolis, 2018.
   - Bordin, Andréa Sabedra. Framework baseado em conhecimento para análise de rede de colaboração científica. Tese - Programa de Pós-graduação em Engenharia, Gestão e Mídia do Conhecimento, Universidade Federal de Santa Catarina, Florianópolis, 2015.
   - Balancieri, Renato. Um método baseado em ontologias para explicitação de conhecimento derivado da análise de redes sociais de um domínio de aplicação. Tese - Programa de Pós-graduação em Engenharia, Gestão e Mídia do Conhecimento, Universidade Federal de Santa Catarina, Florianópolis, 2010.

2) Aprendizado de máquina

- Blanck, Henrique Lopez. Capital de risco e startups: modelo de suporte na tomada de decisão com aprendizado de máquina. Dissertação - Programa de Pós-graduação em Engenharia, Gestão e Mídia do Conhecimento, Universidade Federal de Santa Catarina, Florianópolis, 2020.
- Spillere de Souza, Luiz Fernando. Modelo de mineração de ideias utilizando técnicas de engenharia do conhecimento. Tese - Programa de Pós-graduação em Engenharia, Gestão e Mídia do Conhecimento, Universidade Federal de Santa Catarina, Florianópolis, 2022.

3) Classificação de documento

- Woszezenk, Cristiane Raquel. Modelo para descoberta de conhecimento baseado em associação semântica e temporal entre elementos textuais. Tese - Programa de Pós-graduação em Engenharia, Gestão e Mídia do Conhecimento, Universidade Federal de Santa Catarina, Florianópolis, 2016.
- Adolfo, Luciane Baratto. Uma ontologia de apoio à classificação de processos judiciais. Dissertação - Programa de Pós-graduação em Engenharia, Gestão e Mídia do Conhecimento, Universidade Federal de Santa Catarina, Florianópolis, 2013.

4) Classificação de texto

- Spillere de Souza, Luiz Fernando. Modelo de mineração de ideias utilizando técnicas de engenharia do conhecimento. Tese - Programa de Pós-graduação em Engenharia, Gestão e Mídia do Conhecimento, Universidade Federal de Santa Catarina, Florianópolis, 2022.
- Ceci, Flavio. Um modelo baseado em casos e ontologia para apoio à tarefa intensiva em conhecimento de classificação com foco na análise de sentimento. Tese - Programa de Pós-graduação em Engenharia, Gestão e Mídia do Conhecimento, Universidade Federal de Santa Catarina, Florianópolis, 2015.

5) Descoberta de conhecimento

- Welter, Márcio. Método de identificação de padrões em discurso político a partir da descoberta de conhecimento. Dissertação - Programa de Pós-graduação em

Engenharia, Gestão e Mídia do Conhecimento, Universidade Federal de Santa Catarina, Florianópolis, 2021.

- Ribeiro, Alessandro Costa. Modelo de reconhecimento de padrões em ideias usando técnicas de descoberta de conhecimento em textos. Dissertação - Programa de Pós-graduação em Engenharia, Gestão e Mídia do Conhecimento, Universidade Federal de Santa Catarina, Florianópolis, 2018.
- Woszezenk, Cristiane Raquel. Modelo para descoberta de conhecimento baseado em associação semântica e temporal entre elementos textuais. Tese - Programa de Pós-graduação em Engenharia, Gestão e Mídia do Conhecimento, Universidade Federal de Santa Catarina, Florianópolis, 2016.
- Bovo, Alessandro Botelho. Um modelo de descoberta de conhecimento inerente à evolução temporal dos relacionamentos entre elementos textuais. Tese - Programa de Pós-graduação em Engenharia, Gestão e Mídia do Conhecimento, Universidade Federal de Santa Catarina, Florianópolis, 2011.

6) Mineração de dados

- Sérgio, Marina Carradore. Modelo de avaliação de potenciais ideias alinhadas ao contexto organizacional. Tese - Programa de Pós-graduação em Engenharia, Gestão e Mídia do Conhecimento, Universidade Federal de Santa Catarina, Florianópolis, 2020.
- Koerich, Guilherme Henrique. Conhecimento da marca gastronômica de Florianópolis na mídia Turística com a Chancela UNESCO de Cidade Criativa. Dissertação - Programa de Pós-graduação em Engenharia, Gestão e Mídia do Conhecimento, Universidade Federal de Santa Catarina, Florianópolis, 2020.

7) Mineração de texto

- Trauer, Eduardo. k-SCAS: Framework do Sistema de Agronegócios de Cafés Especiais orientado ao conhecimento. Tese - Programa de Pós-graduação em Engenharia, Gestão e Mídia do Conhecimento, Universidade Federal de Santa Catarina, Florianópolis, 2021.
- Sérgio, Marina Carradore. Modelo de avaliação de potenciais ideias alinhadas ao contexto organizacional. Tese - Programa de Pós-graduação em Engenharia,

Gestão e Mídia do Conhecimento, Universidade Federal de Santa Catarina, Florianópolis, 2020.

- Muniz, Emerson Cleister Lima. Gestão do conhecimento do cliente e destinos turísticos inteligentes: um framework para a gestão inteligente da experiência turística – SMARTUR. Tese - Programa de Pós-graduação em Engenharia, Gestão e Mídia do Conhecimento, Universidade Federal de Santa Catarina, Florianópolis, 2020.

A referência Woszezenk (2013) encontra-se em duas áreas (classificação de documentos e descoberta de conhecimento), enquanto a referência Sérgio (2020) encontra-se em Mineração de Dados (do inglês *Data Mining* – DM) e mineração de texto. Spillere (2022) encontra-se em classificação de texto e aprendizado de máquina.

Dessa forma, diante das referências apresentadas, a presente tese é aderente ao PPGEGC, pois tem o objetivo de propor um modelo voltado à classificação de patentes a partir de fontes de dados não estruturados na forma de texto, levando em conta aspectos de ordenação de subclasses e representação de conhecimento.

1.8 ESTRUTURA DO TRABALHO

O presente trabalho está estruturado em seis capítulos, como se segue:

- o primeiro capítulo apresenta o tema, os objetivos e as delimitações desta pesquisa, juntamente com a aderência ao PPGEGC;
- o segundo capítulo é formado pelo referencial teórico, abordando os principais assuntos relacionados à pesquisa, como análise de patentes, processamento de linguagem natural, representação do conhecimento, aprendizado profundo e trabalhos correlatos;
- o terceiro capítulo descreve a metodologia de pesquisa adotada no desenvolvimento deste trabalho e a revisão da literatura;
- o quarto capítulo apresenta o modelo proposto com seus componentes, bem como um exemplo de instanciação;
- o quinto capítulo é composto pela análise e a discussão dos resultados obtidos por diversas instanciações do modelo em cenários de estudo;
- Por fim, o sexto capítulo traz as considerações finais do trabalho assim como enfatiza possíveis evoluções do modelo proposto e pesquisas futuras.

2 FUNDAMENTAÇÃO TEÓRICA

Este capítulo apresenta a fundamentação teórica envolvida na elaboração desta tese, contendo os temas análise de patentes, processamento de linguagem natural, representação do conhecimento, aprendizado profundo e trabalhos relacionados, como demonstram as seções a seguir.

2.1 ANÁLISE DE PATENTES

À medida que as economias modernas passam a ser intensivas em conhecimento, a Propriedade Intelectual (PI) (do inglês *Intellectual Property* - IP) tem se tornado um meio essencial para as empresas de base tecnológica buscarem vantagem competitiva e garantirem relacionamentos com os clientes. A IP é considerada um ativo intangível, com valor econômico derivado de atividades criativas humanas, e geralmente se enquadra em esquemas de proteção, como patentes, direitos autorais e marcas e registros (Ko *et al*., 2019).

Segundo a WIPO (2021b), a PI refere-se às criações da mente, como invenções, obras literárias e artísticas, designs, símbolos, nomes e imagens usados no comércio. Constitui-se em um meio de proteção ao conhecimento humano dividido em duas categorias: 1) propriedade industrial, que abrange patentes, modelos de utilidade, marcas registradas e desenhos industriais; e 2) indicações geográficas de procedência e direitos autorais, que contemplam obras literárias e artísticas (como romances, poemas, peças teatrais e filmes), composições musicais, produções artísticas (como desenhos, pinturas, fotografias e esculturas) e projetos arquitetônicos.

Neste trabalho, as patentes são selecionadas como objeto de estudo. Uma patente é o direito exclusivo concedido a uma invenção, podendo ser um produto ou processo, geralmente protegido por um período de 20 anos. Dessa forma, o proprietário da patente protege sua invenção da exploração comercial por terceiros (WIPO, 2021b)[3]. O registro e a concessão de uma patente são realizados por um escritório oficial de patentes instituído em cada país ou região.

A apresentação do pedido de registro da patente deve ser realizada por órgão oficial do país ou região, responsável pela análise dos pedidos de patentes. No Brasil, o Instituto Nacional da Propriedade Industrial (INPI), criado em 1970, é uma autarquia federal

[3] Disponível em: www.wipo.int/patents/en. Acesso em: 23 set. 2021.

> responsável por registros de marcas; concessão de patentes; registros de programas de computador; registros de desenho industrial; registros de indicações geográficas; registros de topografia de circuitos integrados; e averbação de contratos de transferência de tecnologia e de franquia empresarial (INPI, 2018, p. 5).

No Quadro 3, são apresentados alguns dos principais escritórios de registro de patentes espalhados pelo mundo.

Quadro 3 – Escritórios de propriedade intelectual

País/Região	Escritório
China	China National Intellectual Property Administration (CNIPA)
Coreia	Korean Intellectual Property Office (KIPO)
EPO	European Patent Office (EPO)
Estados Unidos	United States Patent and Trademark Office (USPTO)
Japão	Japan Patent Office (JPO)
Brasil	Instituto Nacional da Propriedade Industrial (INPI)

Fonte: WIPO (2021)[4]

Um examinador de patentes analisa os pedidos de patentes para determinar se uma dada patente pode ou não ser concedida, além de pesquisar todo o conhecimento tecnológico, também chamado de técnica anterior, para assegurar que a invenção seja nova e única. O examinador também deve revisar os pedidos de patentes para garantir que todos os requisitos de um pedido sejam atendidos e comunicar ao proponente da patente caso alguma norma não esteja de acordo com os requisitos exigidos de patenteabilidade (USPTO, 2021)[5].

O examinador realiza uma pesquisa da técnica anterior e avalia se a invenção definida nas reivindicações é nova e inventiva em relação à técnica anterior encontrada. A esse respeito, o examinador cita principalmente patentes publicadas ou publicações de patentes mais antigas do que livros ou artigos de conferências como técnica anterior (Krestel *et al.*, 2021). Se elementos da invenção reivindicada forem encontrados em uma única referência da técnica anterior, o pedido deixa de ser novidade.

Nas últimas duas décadas ocorreram desenvolvimentos substanciais no campo da análise de patentes, que se caracteriza como a ciência de analisar grandes quantidades de informações de propriedade intelectual em relação a outras fontes de dados para descobrir relacionamentos e tendências (Aristodemou; Tietze, 2018).

Para Wu *et al.* (2016), a análise de patentes é um conjunto de técnicas e ferramentas visuais que determinam a análise de tendências e padrões de inovação tecnológica em um domínio específico baseado em estatísticas de patentes.

[4] Disponível em: https://www.wipo.int/directory/en/urls.jsp. Acesso em: 10 out. 2021.
[5] Disponível em: https://www.uspto.gov/jobs/become-patent-examiner. Acesso em: 10 out. 2021.

A análise de patentes envolve uma série de etapas, incluindo a extração de patentes de bancos de dados de patentes, a extração de informações das patentes e a análise das informações extraídas para inferir as conclusões lógicas. Os tipos de conteúdo de patentes são variados, uma vez que os dados podem ser estruturados e não estruturados (Abbas; Zhang; Khan, 2014). Os dados de patentes estruturados contêm algumas informações, como o inventor da patente, o responsável por ela e informações de citação. Já os dados de patentes não estruturadas compreendem o texto narrativo, incluindo o título da patente, o resumo, as reivindicações e a descrição (Tseng; Lin; Lin, 2007).

Segundo Moehrle *et al.* (2010), as etapas para o processo da análise dos documentos de patentes passam pela preparação dos dados para um acesso eficiente, a análise dos dados por meio de diferentes métodos e a utilização dos resultados da análise para a tomada de decisão. Dessa forma, em um contexto de negócio, a análise dos documentos de patentes consiste em três etapas principais: 1) pré-processamento, 2) análise de patentes e 3) conhecimento descoberto.

Na etapa de pré-processamento, os dados são coletados dos documentos de patentes com o objetivo de fornecer informações com qualidade, exatidão e integridade. A etapa de análise de patentes consiste na análise dos dados extraídos na etapa de pré-processamento. Isso ocorre através de diferentes métodos e técnicas para classificar, agrupar e identificar percepções significativas a partir das informações. Já a etapa de conhecimento descoberto, os resultados e as informações da etapa da análise de patentes são visualizados e avaliados para apoiar a tomada de decisões estratégicas (Aristodemou; Tietze, 2018; Moehrle *et al.*, 2010).

Como pode ser visto, esses processos exigem que os analistas tenham certo grau de experiência em recuperação de informações, tecnologias específicas de domínio e inteligência de negócios. Assim, encontrar um examinador[6] que preencha esses requisitos torna-se uma tarefa difícil, além de requerer um processo dispendioso de treinamento. Além disso, os documentos de patentes costumam ser longos e ricos em terminologia técnica e jurídica, o que consome tempo considerável para ler e analisar, mesmo para examinadores experientes (Tseng; Lin; Lin, 2007).

[6] Um examinador é um especialista em algum campo de invenção (Risch; Krestel, 2018).

2.1.1 Tarefas de análise de patentes

Além dos conceitos sobre análise de patentes vistos na seção anterior, essa ciência de análise consiste em resolver algumas tarefas que dependem da intenção do usuário, empresa, inventor ou pesquisadores. Segundo Bonino, Ciaramella e Corno (2010), os usuários querem obter informações essenciais e complementares, empresas e inventores têm interesse em pesquisar se a invenção é realmente uma novidade, enquanto os pesquisadores desejam encontrar informações sobre patentes para evitar a duplicação de soluções já cobertas por patentes e/ou reutilizar livremente patentes expiradas.

Dessa forma, os dados de patentes se tornam de vital importância para escritórios e organizações requerentes de patentes, pois, segundo Abbas, Zhang e Khan (2014), auxiliam os tomadores de decisão de várias maneiras para atender diferentes tarefas.

Na literatura, alguns autores definem tarefas que podem ser executadas com a análise de patentes. Bonino, Ciaramella e Corno (2010) subdividem as tarefas em pesquisa, análise e monitoramento de patentes. A pesquisa diz respeito à recuperação de informações relevantes sobre patentes, geralmente utilizadas para procedimentos de pedidos de patentes, com o intuito de reduzir o risco de violação de patentes. A análise é uma atividade muito ampla; torna-se necessário analisar uma única patente ou um portfólio de patentes, de acordo com o objetivo específico da decisão estratégica a ser tomada. Já o monitoramento está relacionado com a necessidade de se monitorar tanto o cenário tecnológico, para avaliar a evolução das trajetórias tecnológicas, quanto o comportamento inovador dos concorrentes, a fim de identificar seus perfis tecnológicos (Baglieri; Cesaroni, 2013).

Abbas, Zhang e Khan (2014) descrevem algumas tarefas executadas com as informações da análise de patentes, entre as quais estão as seguintes: analisar e prever tendências tecnológicas; conduzir planejamento estratégico de tecnologia; detectar violação de patentes; determinar a qualidade de patentes; e indicar patentes mais promissoras.

Já Krestel *et al.* (2021) fizeram uma revisão na literatura das tarefas mais populares para a análise de patentes, classificadas em:

- Suporte: compreende o pré-processamento, a extração de informações para análise posterior ou a tradução de patentes para outros idiomas;
- Classificação de patentes: os documentos de patentes são categorizados hierarquicamente com base no campo de invenção;

- Recuperação de patentes: compreende a busca de técnica anterior, criação e avaliação automatizada de cenários de patentes, busca de violação, busca de liberdade de operação e recuperação de passagem[7];
- Avaliação de patentes e previsão do valor de mercado: pesquisa inovadora em que o conteúdo e os detalhes bibliográficos das patentes são analisados com certos protocolos humanos para avaliar a qualidade dos pedidos de patentes. Essa análise é posteriormente usada para adicionar valor de mercado;
- Previsão de tecnologia: as patentes são empregadas para avaliar um cenário tecnológico com o intuito de auxiliar pesquisadores na identificação de novas tecnologias ou tecnologias do momento;
- Geração de textos de patentes: a estrutura e os estilos incorporados aos documentos de patentes publicados são usados para automatizar o processo de redação de reivindicações de patentes;
- Análise de litígios: processo legal em que patentes potenciais levam a uma disputa ou litígio entre quaisquer duas empresas que acabam por proibir o desenvolvimento de estratégias de negócios;
- Tarefas de visão computacional (do inglês: *Computer Vision* - CV): trabalham com figuras e desenhos de documentos de patentes em vez de texto.

2.1.2 Escritórios de patentes

Os escritórios de patentes são organizações governamentais ou não governamentais que supervisionam a emissão de patentes em um país ou região. Como as patentes são territoriais, ou seja, oferecem proteção somente na jurisdição onde o pedido foi depositado e a patente concedida, determinado escritório de patentes examina cada pedido para ver se está de acordo com os requisitos exigidos na legislação. Além disso, a invenção deve se enquadrar em um assunto patenteável nas jurisdições onde foi depositado (WIPO, 2019).

No Quadro 4 são descritos os cinco principais escritórios de patentes com maior número de pedidos em 2019 segundo a WIPO (2020).

No ano de 2019 a Administração Nacional de Propriedade Intelectual da República Popular da China (CNIPA) recebeu 1,4 milhão de pedidos de patentes, número superior ao

[7] Um sistema de recuperação de passagens pode ser definido como um tipo especializado de aplicação de recuperação da informação que recupera passagens relevantes (partes de textos) em vez de fornecer um conjunto de documentos classificados (Buscaldi *et al*., 2010).

dobro do recebido pelo USPTO e mais que a soma dos pedidos de patentes dos escritórios USPTO, JPO, KIKO e EPO. Os cinco escritórios de patentes juntos respondem por 84,7% do total mundial de pedidos de patentes em 2019 (WIPO, 2020). Em 2020, apesar do aumento no número total de pedidos de patentes, somente os escritórios da China e da Coreia tiveram aumento no número de pedidos de patentes (WIPO, 2021a). Em 2021, apenas o escritório USPTO não teve aumento no número de depósitos de patentes (WIPO, 2022a).

Quadro 4 – Principais escritórios de patentes

Escritório de patente	**Sigla**	**País**	**Pedidos de patentes**		
			2019	**2020**	**2021**
The National Intellectual Property Administration of the People's Republic of China	CNIPA	China	1.400.000	1.500.000	1.590.000
United States Patent and Trademark Office	USPTO	EUA	621.453	597.172	591.473
Japan Patent Office	JPO	Japão	307.969	288.472	289.200
Korean Intellectual Property Office	KIPO	Coreia	218.975	226.759	237.998
European Patent Office	EPO	Europa	181.479	180.346	188.778
Instituto Nacional da Propriedade Industrial[8]	INPI	Brasil	28.318	27.091	26.921

Fonte: adaptado de World Intellectual Property Indicators (2023)

Apesar de o número do pedido de patentes ter diminuído em 2019 em relação a 2018, voltou a crescer nos anos de 2020 e 2021. Os escritórios de patentes têm a difícil tarefa de analisar os documentos de patentes. É crescente a necessidade de automação por parte desses escritórios para atender a demanda dos pedidos de patentes. Nesse sentido, métodos e técnicas recentes de representação do conhecimento, aprendizado de máquina e aprendizado profundo podem contribuir com a automatização das tarefas de análise de patentes. A WIPO e os escritórios de patentes como USPTO e EPO já reconheceram o potencial de tais técnicas e vêm desenvolvendo pesquisas para promover a aplicação em tarefas como classificação de patente, pesquisa da arte anterior, análise de dados, exame de patentes e análise de imagens (Krestel *et al*., 2021).

Segundo Krestel *et al.* (2021), a pesquisa sobre aprendizado profundo para o domínio de patentes é, portanto, não apenas conduzida pela academia e a indústria, mas é ativamente alimentada pelos principais escritórios de patentes, fornecendo conjuntos de dados de referência, organizando desafios e promovendo workshops.

[8] Os dados do INPI foram extraídos do Boletim Mensal de Propriedade Industrial de março de 2022. Disponível em: https://www.gov.br/inpi/pt-br/central-de-conteudo/estatisticas/arquivos/publicacoes/boletim-mensal-de-propriedade-industrial-marco-de-2022.pdf. Acesso em: 26 abr. 2023.

2.1.3 Métodos e técnicas recentes

O crescente aumento no número de pedidos de patentes vem criando desafios consideráveis para todo o sistema de patentes e também para os usuários de informações dessa fonte. Os examinadores de patentes têm papel fundamental no andamento das concessões ou não dos pedidos de patentes. Como visto na seção anterior, os principais escritórios de patentes estão envolvidos em desenvolver ferramentas para automatizar processos da análise, em especial com a utilização de AI e ML (do inglês *Machine Learning* - ML). Segundo Li *et al.* (2018), a eficiência e a precisão em análise de patentes serão melhoradas com a aplicação da AI para acelerar, por exemplo, a classificação de patentes, que é uma das tarefas essenciais durante a etapa de avaliação de documentos para a concessão da patente.

O ML, DL (do inglês *Deep Learning* - DL) e as ANN (do inglês *Artificial Neural Network* - ANN) estão entre as técnicas de AI predominantes na classificação de patentes. No Quadro 5, percebe-se que o aprendizado de máquina com algoritmos mais tradicionais para a classificação de patentes ainda é utilizado. Entre os principais algoritmos destacam-se:

- Naïve Bayes (Cassidy, 2020; Liu *et al*., 2020; Naik; Brunda; Seema, 2020; Shahid *et al*., 2020; Xiao; Wang; Liu, 2018);
- K-Nearest Neighbors (k-NN) (De Clercq *et al*., 2019; Shahid *et al*., 2020; Xiao; Wang; Zuo, 2018; Yücesoy Kahraman; Dereli; Durmuşoğlu, 2018);
- Support Vector Machine (SVM): (Frerich *et al*., 2021; Jafery *et al*., 2019; Shahid *et al*., 2020; Yücesoy Kahraman; Dereli; Durmuşoğlu, 2018; Yun; Geum, 2020);
- Decision tree (DT): (De Clercq *et al*., 2019; Naik; Brunda; Seema, 2020);
- Random Forest (RF): (De Clercq *et al*., 2019);
- Latent Dirichlet allocation (LDA): (Yun; Geum, 2020);
- Singular Value Decomposition (SVD): (Huang; Tan, 2020).

Já os autores Aroyehun *et al.* (2021), Shahid *et al.* (2020) e Trappey, Trappey e Hsieh (2021) combinam a técnica de ML com ANN para classificação de documentos de patentes.

Geralmente a representação do documento nos algoritmos de ML se utiliza de listas de palavras/termos com suas respectivas frequências, *n*-gramas ou frases, dependendo da frequência de ocorrência e/ou coocorrência de termos ou palavras. Já o aprendizado profundo e as redes neurais baseiam-se na semântica e na ordem das palavras no documento (Aroyehun *et al.*, 2021).

Os algoritmos mais utilizados de ML e ANN encontrados na literatura para classificação de patentes são:

- Recurrent Neural Network (RNN): (Grawe; Martins; Bonfante, 2017; Risch; Krestel, 2019);
- Longshort-Term-Memory (LSTM): (Grawe; Martins; Bonfante, 2017; Hu *et al*., 2018; Huang *et al*., 2020; Sofean, 2021); e
- Convolutional Neural Network (CNN): (Abdelgawad *et al*., 2019; Bai; Shim; Park, 2020; Chung; Sohn, 2020; Hu *et al*., 2018; Li *et al*., 2018; Lu *et al*., 2019; Zhu *et al*., 2020).

2.1.4 Fontes de informação

Os principais escritórios de patentes – CNIPA, USPTO, JPO, KIPO e EPO – e a própria WIPO disponibilizam ferramentas para a consulta de patentes. Esses escritórios emitem semanalmente um conjunto de dados com os pedidos de patentes que podem ser coletados, geralmente no formato XML[9]. Esse conjunto de dados possui informações sobre os documentos de patentes, os quais são estruturados em seções, incluindo título, resumo, histórico da invenção, descrição e reivindicações.

A seção de reivindicações, em complemento às demais, é entendida como relevante, pois descreve o escopo da proteção buscada pelo inventor e, portanto, codifica o valor real da patente. Os documentos de patentes são extensos, com terminologia bastante complexa e específica de cada domínio. Podem ser compostos por texto, imagens, fluxogramas, fórmulas, com um rico conjunto de metadados e informações bibliográficas (por exemplo, códigos de classificação, citações, inventores, cessionário, datas de arquivamento/publicação, endereços, examinadores) (Shalaby; Zadrozny, 2019).

A utilização do título e resumo de patentes é comum na composição de conjuntos de dados para tarefas de NLP (do inglês *Natural Language Processing* - NLP) aplicadas ao domínio de patentes. Essas seções contêm descrições concisas e gerais das invenções, o que as torna adequadas para representar o conteúdo das patentes na maioria dos casos. Elas equilibram a relevância do texto com viabilidade computacional. Porém, dependendo da complexidade da tarefa, pode ser necessário incorporar outras seções, como a descrição técnica, para obter melhor desempenho. Portanto, não há uma regra definitiva sobre quais seções utilizar. O ideal

[9] Extensible Markup Language.

é testar e avaliar, para cada caso, a combinação mais apropriada entre precisão, custo computacional e disponibilidade dos dados (Ruijie *et al*., 2021).

Além dos conjuntos de dados disponibilizados pelos escritórios de patentes, algumas instituições como o Conference and Labs of the Evaluation Forum - Intelectual Property Task (CLEF-IP[10]) e o NII Testbeds and Community for Information Access Research (NTCIR[11]) fornecem uma coleção de dados para workshops, conferências e competições. O laboratório CLEF-IP reúne documentos da EPO com dados de 2009 a 2013 em três idiomas (inglês, alemão e francês) com o objetivo de estimular a pesquisa na área de recuperação de patentes e organizar um conjunto de dados preparado para realizar testes com os dados de patentes (Piroi; Lupu; Hanbury, 2013).

Já o NTCIR é uma série de workshops de avaliação projetada para aprimorar a pesquisa em tecnologias de acesso à informação, incluindo recuperação de informação, resposta a perguntas, resumo de texto, extração, entre outras. O principal objetivo do NTCIR é fornecer coleções de testes reutilizáveis em grande escala para experimentos, bem como uma infraestrutura de avaliação comum que permita comparações entre sistemas (NTCIR, 2021).

Na revisão integrativa, descrita na seção 3.3, foram identificados os conjuntos de dados utilizados em diversos trabalhos (Quadro 5). Percebe-se que os autores buscaram sites especializados em patentes com conjuntos de dados já disponíveis ou criaram o próprio conjunto de dados a partir de documentos disponibilizados na *web*.

Quadro 5 – Conjunto de dados na classificação de patentes

(continua)

id	Autor	Conjunto de dados
1	(Liu *et al.*, 2020)	*Crawler* (www.baiten.cn)
2	(Huang; Tan, 2020)	Recuperação de patentes no ano de 2000, utilizando ontologia e computação em nuvem
3	(Yun; Geum, 2020)	USPTO
4	(Jafery *et al*., 2019)	MyIPO
5	(De Clercq *et al*., 2019)	Banco de dados PatentsView SQL hospedado no Google BigQuery
6	(Frerich *et al*., 2021)	Otimizados por especialistas e PATSTAT
7	(Cassidy, 2020)	WPI *test collection*
8	(Yücesoy Kahraman; Dereli; Durmuşoğlu, 2018)	WIPO-alpha
9	(Xiao; Wang; Liu, 2018)	Patentes da área de segurança extraídas de sites
10	(Xiao; Wang; Zuo, 2018)	Wikipedia chinesa

(conclusão)

[10] Disponível em: http://https://clef2022.clef-initiative.eu. Acesso em: 2 maio 2023.

[11] Disponível em: http://research.nii.ac.jp/ntcir/index-en.html. Acesso em: 2 maio 2023.

id	Autor	Conjunto de dados
11	(Shahid *et al*., 2020)	USPTO
12	(Trappey; Trappey; Hsieh, 2021)	USPTO
13	(Aroyehun *et al*., 2021)	WIPO-alpha, USPTO-2M
14	(Yu *et al*., 2020)	4 conjuntos de dados de patentes baseados em TRIZ[12] de patentes chinesas, coletados da Baiten (https://www.baiten.cn/) (2007 a 2015)
15	(Grawe; Martins; Bonfante, 2017)	USPTO (2006 a 2017)
16	(Roudsari *et al*., 2020)	USPTO-3M
17	(Min, 2021)	Patentes de energia (China)
18	(Risch; Krestel, 2019)	WIPO, USPTO-2M e CLEF-IP
19	(Chung; Sohn, 2020)	USPTO (2000-2015)
20	(Bai; Shim; Park, 2020)	USPD e AAPD (1976-2020)
21	(Abdelgawad *et al*., 2019)	Wipo-Alpha e Pat16
22	(Li *et al*., 2018)	CLEF-IP e USPTO-2M
23	(Sofean, 2021)	EPO e WIPO
24	(Naik; Brunda; Seema, 2020)	www.baiten.cn
25	(Lu *et al*., 2019)	Corpus de patente extraídos de site
26	(Huang *et al*., 2020)	3 conjuntos de patentes chinesas foram criados
27	(Zhu *et al*., 2020)	Patentes chinesas e CNKI
28	(Hu *et al*., 2018)	CLEF-IP 2011 (M-CLEF)
29	(Haghighian Roudsari *et al*., 2022)	USPTO-2 e CLEF-IP 2011
30	(Jiang *et al*., 2022)	USPTO
31	(Choi *et al*., 2022)	KISTA (Korea Intellectual Property Service Center)
32	(Liu *et al*., 2021)	Patentes de logística da China
33	(Lo; Chu, 2021)	USPTO (2016 a 2021)

Fonte: elaborado pelo autor (2023)

Dos 33 artigos levantados na literatura, 8 autores utilizaram conjuntos de dados de patentes chinesas. Os autores Liu *et al.* (2020) e Naik, Brunda e Seema (2020) extraíram os dados do site da Baiten®[13], que concentra publicações de patentes chinesas. Já Huang *et al.* (2020) e Yu *et al.* (2020) também usaram conjunto de dados de patentes chinesas, mas não especificaram de onde extraíram esses dados.

Xiao, Wang e Zuo (2018) utilizaram a Wikipedia chinesa como conjunto de dados, enquanto Min (2021) usou dados de patentes chinesas sobre energia, não especificando o local da extração. Liu *et al.* (2021) fizeram uso de patentes de logística da China.

Por fim, Zhu *et al.* (2020) utilizaram dois conjuntos de dados: (1) conjunto de documentos de patentes chinesas em pequena escala, com pouco mais de 3.000 patentes

[12] TRIZ (do inglês *Theory of Inventive Problem Solving* ou Teoria da Resolução Inventiva de Problemas) tem como objetivo ajudar a encontrar soluções criativas e eficientes para problemas complexos (Park; Ree; Kim, 2012).

[13] Disponível em: http://www.baiten.cn. Acesso em: 4 jul. 2021.

chinesas originais; e (2) dados rastreados do banco de dados de patentes publicado da Chinese National Knowledge Infrastructure[14] (CNKI).

A maioria dos artigos (14) manusearam conjuntos de dados de patentes dos Estados Unidos, extraídos principalmente do USPTO ou combinados com outros conjuntos de dados. Vários trabalhos empregaram somente conjuntos de dados extraídos da USPTO® (Chung; Sohn, 2020; Grawe; Martins; Bonfante, 2017; Jiang *et al.*, 2022; Lo; Chu, 2021; Roudsari *et al.*, 2020; Shahid *et al.*, 2020; Trappey; Trappey; Hsieh, 2021; Yun; Geum, 2020). Já Aroyehun *et al.* (2021) combinaram os dados da USPTO juntamente com o conjunto de dados da WIPO. Por sua vez, Risch e Krestel (2019), além dos dados da USPTO e WIPO, também utilizaram o conjunto de dados da CLEF-IP no estudo, e Haghighian Roudsari *et al.* (2022) combinaram os dados da USPTO com o conjunto de dados da CLEF-IP.

Bai, Shim e Park (2020) fizeram uso do conjunto de dados patenteado dos EUA (USPD)[15], também fornecido pela USPTO, bem como do conjunto de dados da Arxiv Academic Paper Dataset (AAPD)[16], que contém o resumo de 55.840 artigos na área da Ciência da Computação. Cassidy (2020) utilizou patentes dos EUA extraídas da World Patent Information (WPI). Li *et al.* (2018) combinaram os dados da USPTO com os dados da CLEF-IP como base para o seu estudo. Já De Clercq *et al.* (2019) utilizaram o banco de dados da *PatentsView* na pesquisa.

Yücesoy Kahraman, Dereli e Durmuşoğlu (2018) valeram-se do conjunto fornecido pela WIPO, enquanto Abdelgawad *et al.* (2019) combinaram os dados da WIPO e da CLEF-IP para compará-los com um conjunto de dados chamado de Pat, que consiste em uma coleção não divulgada de 1,05 milhão de patentes inglesas depositadas em um escritório de patentes. Sofean (2021) combinou os conjuntos de dados da WIPO com a EPO para a realização dos seus estudos.

O conjunto de dados da Intellectual Property Corporation os Malaysia (MyIPO) foi utilizado pelos autores Jafery *et al.* (2019), enquanto os autores Hu *et al.* (2018) realizaram os estudos com o conjunto de dados disponibilizado pela CLEF-IP. Já Choi *et al.* (2022) fizeram uso do conjunto de dados do Korea Intellectual Property Service Center (KISTA[17]). O KISTA é uma organização afiliada do Escritório Coreano de Propriedade Intelectual (KIPO), responsável pela análise de patentes.

[14] Disponível em: http://cnki.scstl.org/kns55/brief/result.aspx?dbPrefix=SCPD. Acesso em: 4 jul. 2021.
[15] Disponível em: http://www.patentsview.org/download/. Acesso em: 4 jul. 2021.
[16] Disponível em: https://github.com/lancopku/SGM. Acesso em: 10 jul. 2021.
[17] Disponível em: https://www.kista.re.kr/?lang=EN. Acesso em: 10 jul. 2021.

Os demais autores não especificaram a origem dos dados extraídos. Por exemplo, Xiao, Wang e Liu (2018) extraíram patentes da área de segurança; Huang e Tan (2020) utilizaram patentes de *cloud computing* como base de dados; Frerich *et al.* (2021) coletaram patentes da base de dados da PATSTAT; Lu *et al.* (2019) coletaram um *corpus* de patentes sem especificar o site utilizado para tal.

Na revisão da literatura, percebe-se que os grandes escritórios de patentes USPTO e EPO, o site Baiten, especializado em serviços de propriedade intelectual, e a WIPO disponibilizam uma coleção de dados prontos para pesquisa, ou então o pesquisador pode ir até o site do escritório e extrair os dados para construir o seu próprio conjunto de dados, de acordo com a necessidade e o contexto da pesquisa.

2.2 PROCESSAMENTO DE LINGUAGEM NATURAL

De acordo com Allhyari *et al.* (2017), o Processamento de Linguagem Natural (do inglês *Natural Language Processing* - NLP) é um subcampo da ciência da computação, IA e linguística que visa a compreensão da linguagem natural usando computadores, sendo muito utilizado para mineração de texto. Lauriola, Lavelli e Aiolli (2022) definem NLP como um ramo da AI repleto de tarefas complexas, sofisticadas e desafiadoras relacionadas à linguagem, como a tradução automática, respostas a perguntas, resumo de textos e assim por diante. O NLP envolve o projeto e a implementação de modelos, sistemas e algoritmos para resolver problemas práticos de compreensão de linguagens humanas. Ou seja, a tecnologia tem como objetivo extrair as informações e percepções contidas nos documentos, bem como categorizar e organizar os próprios documentos.

O NLP representa parte indispensável do avanço da era da AI e vem sendo utilizado no domínio da interface/interação homem-computador (do inglês *Human-Computer Interaction* - HCI) para todos os aplicativos de software de última geração. O NLP permite estabelecer interfaces entre máquinas e humanos, por meio do entendimento de linguagens humanas que envolvam diálogos (Chao *et al.*, 2021).

A linguagem natural é uma fonte valiosa e rica de informações para muitas aplicações. Ainda assim, é discreta e esparsa e, como tal, uma fonte de dados desafiadora. Para que o texto seja utilizável como dado de entrada, deve-se primeiro transformá-lo em uma representação adequada, que geralmente é um vetor dos recursos do texto, portanto um vetor de números. As representações vetoriais de texto podem ser construídas de muitas maneiras diferentes (Babić; Martinčić-Ipšić; MeštroviĆ, 2020).

De acordo com o foco das tarefas de NLP, os métodos podem ser categorizados em dois tipos. O primeiro se refere à análise de sintaxe, estando relacionado com a compreensão de estruturas de palavras, frases e documentos. As tarefas típicas incluem segmentação morfológica, segmentação de palavras, marcação de classe gramatical e análise sintática. O segundo refere-se à análise semântica. Tem como objetivo compreender os significados das palavras, frases e suas combinações. As tarefas típicas incluem Reconhecimento de Entidade Nomeadas (do inglês *Named Entity Recognition* - NER), Análise de Sentimento (do inglês *Sentiment Analysis* - SA), Tradução Automática (*Machine Translation*) e resposta a perguntas (do inglês *Question Answering* - QA).

De modo geral, antes de qualquer processamento de linguagem natural de um texto, este deve ser normalizado (Jurafsky; Martin, 2021). Entre as funções importantes no tratamento do texto citam-se:

- Tokenização: separa o texto em palavras;
- Frequência/contagem de *tokens*: realiza a contagem de ocorrência de *tokens*[18];
- Remoção de *stopwords*: remove *tokens* irrelevantes;
- *N*-gramas: sequência de palavras onde *n* indica o número de palavras;
- Um bigrama é a sequência de duas palavras, trigrama uma sequência de três palavras e assim por diante;
- *Stemming*: consiste em reduzir a palavra ao seu radical;
- Lematização: permite reduzir a palavra à sua forma canônica, levando em conta sua classe gramatical;
- Etiquetagem: classifica cada ocorrência de uma palavra em uma frase como, por exemplo, um substantivo, adjetivo ou verbo;
- NER: busca extrair e classificar as entidades mencionadas em um texto escrito em linguagem natural.

Parte dessas funções está disponível na biblioteca NLTK®[19], muito utilizada em linguagem natural no desenvolvimento de softwares em linguagem Python®.

Para exemplificar as funções do processamento de linguagem natural, é apresentada uma patente obtida a partir do Google Patents® de número US08001811[20], que tem como título

[18] *Token* pode ser entendido como uma unidade linguística indivisível em determinado domínio, por exemplo, uma palavra ou um URL.
[19] Disponível em: https://www.nltk.org. Acesso em: 28 jul. 2021.
[20] Disponível em: https://patents.google.com/patent/EP1995367A2/en. Acesso em: 28 jul. 2021.

"*Washing machine having water softening device*". Na Figura 2, consta o *abstract* dessa patente.

Figura 2 – *Abstract* da patente US08001811

A washing machine having a water softening device which improves the solubility of a detergent and a water sof tening performance concurrently. The washing machine includes a tub; a water supply device for supplying wate r to the tub; a detergent supply device for supplying a detergent to the tub; and a water softening device for softe ning the water. The water softening device is disposed such that the water supplied from the water supply device can be mixed with the detergent supplied from the detergent supply device and then the water mixed with the det ergent can be supplied to the water softening device.

Fonte: elaborado pelo autor (2023)

O *abstract* da patente descreve sucintamente a invenção, no caso um resumo da patente US08001811. A partir do *abstract*, é aplicada a técnica de tokenização. Na Figura 3, é possível perceber que o texto foi separado em palavras, sendo removidos números e pontuações.

Figura 3 – Tokenização do *abstract* da patente US08001811

['washing', 'machine', 'having', 'water', 'softening', 'device', 'which', 'improves', 'the', 'solubility', 'of', 'detergent', 'a nd', 'water', 'softening', 'performance', 'concurrently', 'The', 'washing', 'machine', 'includes', 'tub', 'water', 'supply', ' device', 'for', 'supplying', 'water', 'to', 'the', 'tub', 'detergent', 'supply', 'device', 'for', 'supplying', 'detergent', 'to', 'the' , 'tub', 'and', 'water', 'softening', 'device', 'for', 'softening', 'the', 'water', 'The', 'water', 'softening', 'device', 'is', 'dispo sed', 'such', 'that', 'the', 'water', 'supplied', 'from', 'the', 'water', 'supply', 'device', 'can', 'be', 'mixed', 'with', 'the', 'det ergent', 'supplied', 'from', 'the', 'detergent', 'supply', 'device', 'and', 'then', 'the', 'water', 'mixed', 'with', 'the', 'deterge nt', 'can', 'be', 'supplied', 'to', 'the', 'water', 'softening', 'device']

Fonte: elaborado pelo autor (2023)

Com o texto transformado em *tokens*, pode-se realizar a frequência/contagem desses *tokens* gerados na etapa anterior (Figura 4), ordenados pelas suas frequências, nesse caso com um total de 92 ocorrências de *tokens*.

Figura 4 – Frequência/contagem de *tokens*

[('water', 11), ('the', 11), ('device', 8), ('softening', 6), ('detergent', 6), ('supply', 4), ('and', 3), ('tub', 3), ('for', 3), ('t o', 3), ('supplied', 3), ('washing', 2), ('machine', 2), ('The', 2), ('supplying', 2), ('from', 2), ('can', 2), ('be', 2), ('mixe d', 2), ('with', 2), ('having', 1), ('which', 1), ('improves', 1), ('solubility', 1), ('of', 1), ('performance', 1), ('concurrent ly', 1), ('includes', 1), ('is', 1), ('disposed', 1), ('such', 1), ('that', 1), ('then', 1)]

Fonte: elaborado pelo autor (2023)

Na Figura 5, apresenta-se a frequência dos *tokens* com a remoção das *stopwords*, resultando num total de 55 ocorrências de *tokens*. Os *tokens* "*the*", "*and*", "*for*", "*to*", "*The*", "*from*", "*can*", "*be*", "*with*", "*having*", "*which*", "*of*", "*is*", "*such*", "*that*" e "*then*" foram removidos da lista.

Figura 5 – Remoção de *stopwords*

[('water', 11), ('device', 8), ('softening', 6), ('detergent', 6), ('supply', 4), ('tub', 3), ('supplied', 3), ('washing', 2), ('machine', 2), ('supplying', 2), ('mixed', 2), ('improves', 1), ('solubility', 1), ('performance', 1), ('concurrently', 1), ('includes', 1), ('disposed', 1)]

Fonte: elaborado pelo autor (2023)

A partir da lista de *tokens*, são criados bigramas para identificar os pares de *tokens* existentes como exemplo de *n-grams*, como mostra a Figura 6. Exemplos de bigramas são: (“*washing*”, “*machine*”), (“*machine*”, “*water*”), (“*water*”, “*softening*”), (“*softening*”, “*device*”), entre outros.

Figura 6 – Bigramas de *tokens*

[('washing', 'machine'), ('machine', 'water'), ('water', 'softening'), ('softening', 'device'), ('device', 'washing'), ('washing', 'machine'), ('machine', 'water'), ('water', 'softening'), ('softening', 'device'), ('device', 'improves'), ('improves', 'solubility'), ('solubility', 'detergent'), ('detergent', 'water'), ('water', 'softening'), ('softening', 'performance'), ('performance', 'concurrently'), ('concurrently', 'washing'), ('washing', 'machine'), ('machine', 'includes'), ('includes', 'tub'), ('tub', 'water'), ('water', 'supply'), ('supply', 'device'), ('device', 'supplying'), ('supplying', 'water'), ('water', 'tub'), ('tub', 'detergent'), ('detergent', 'supply'), ('supply', 'device'), ('device', 'supplying'), ('supplying', 'detergent'), ('detergent', 'tub'), ('tub', 'water'), ('water', 'softening'), ('softening', 'device'), ('device', 'softening'), ('softening', 'water'), ('water', 'water'), ('water', 'softening'), ('softening', 'device'), ('device', 'disposed'), ('disposed', 'water'), ('water', 'supplied'), ('supplied', 'water'), ('water', 'supply'), ('supply', 'device'), ('device', 'mixed'), ('mixed', 'detergent'), ('detergent', 'supplied'), ('supplied', 'detergent'), ('detergent', 'supply'), ('supply', 'device'), ('device', 'water'), ('water', 'mixed'), ('mixed', 'detergent'), ('detergent', 'supplied'), ('supplied', 'water'), ('water', 'softening'), ('softening', 'device'), ('device', 'washing'), ('washing', 'machine'), ('machine', 'water'), ('water', 'softening'), ('softening', 'device')]

Fonte: elaborado pelo autor (2023)

Outra possibilidade de pré-processamento é a aplicação de *stemming*, em que são removidos os sufixos das palavras. O Quadro 6 apresenta algumas palavras retiradas do resumo da patente e o *stemming* correspondente.

Quadro 6 – *Stemming* das palavras

Palavra	***Stemming***
softening	*soften*
washing	*wash*
supplying	*suppli*
supplied	*suppli*
mixed	*mix*
solubility	*solubl*

Fonte: elaborado pelo autor (2023)

Além do *stemming*, a lematização também pode ser utilizada. Possui como tarefa verificar se duas palavras possuem a mesma raiz, apesar das suas diferenças superficiais. No Quadro 7, a lematização foi aplicada ao *abstract* de uma patente. Na segunda palavra do quadro,

percebe-se a conversão da palavra "*having*", que é um verbo de ligação, para o seu verbo-raiz, no caso o verbo "*have*".

Quadro 7 – Lematização das palavras

Palavra	**Lematização**
disposed	*dispose*
having	*have*
improves	*improve*
includes	*include*
mixed	*mix*
softening	*soften*
supplied	*supply*
supplying	*supply*
washing	*wash*

Fonte: elaborado pelo autor (2023)

Por fim, a função de etiquetagem processa uma sequência de palavras e anexa uma *tag* gramatical a cada palavra. Na Figura 7, verifica-se que à palavra "*machine*" foi adicionada a etiqueta "*NN*", o que indica que essa palavra é um substantivo singular ou não contável. As palavras "*washing*" e "*softening*" receberam a etiqueta "*VBG*", indicando um verbo gerúndio ou particípio presente. A palavra "*supplied*" tem a etiqueta "*VBD*", que indica que essa palavra é um verbo no pretérito. E assim, todas as palavras do *abstract* receberam uma etiqueta referente à classe gramatical correspondente.

Figura 7 – Etiquetagem de palavras

```
[('washing', 'VBG'), ('machine', 'NN'), ('water', 'NN'), ('softening', 'VBG'), ('device', 'NN'), ('washing', 'VBG'), ('m
achine', 'NN'), ('water', 'NN'), ('softening', 'VBG'), ('device', 'NN'), ('improves', 'VBZ'), ('solubility', 'JJ'), ('deterge
nt', 'NN'), ('water', 'NN'), ('softening', 'VBG'), ('performance', 'NN'), ('concurrently', 'RB'), ('washing', 'VBG'), ('m
achine', 'NN'), ('includes', 'VBZ'), ('tub', 'JJ'), ('water', 'NN'), ('supply', 'NN'), ('device', 'NN'), ('supplying', 'VBG'),
('water', 'NN'), ('tub', 'JJ'), ('detergent', 'NN'), ('supply', 'NN'), ('device', 'NN'), ('supplying', 'VBG'), ('detergent', 'N
N'), ('tub', 'NN'), ('water', 'NN'), ('softening', 'VBG'), ('device', 'NN'), ('softening', 'VBG'), ('water', 'NN'), ('water',
'NN'), ('softening', 'VBG'), ('device', 'NN'), ('disposed', 'VBD'), ('water', 'NN'), ('supplied', 'VBD'), ('water', 'NN'),
('supply', 'NN'), ('device', 'NN'), ('mixed', 'JJ'), ('detergent', 'NN'), ('supplied', 'VBD'), ('detergent', 'JJ'), ('supply', '
NN'), ('device', 'NN'), ('water', 'NN'), ('mixed', 'JJ'), ('detergent', 'NN'), ('supplied', 'VBD'), ('water', 'NN'), ('softeni
ng', 'VBG'), ('device', 'NN'), ('washing', 'VBG'), ('machine', 'NN'), ('water', 'NN'), ('softening', 'VBG'), ('device', '
NN')]
```

Fonte: elaborado pelo autor (2023)

Na atualidade, o NLP é principalmente um campo orientado por dados que usa cálculos estatísticos e probabilísticos junto com ML. No entanto, durante os últimos anos, houve uma transformação generalizada nesse sentido, e essas abordagens foram em grande parte substituídas, ou pelo menos aprimoradas, por modelos neurais (Otter; Medina; Kalita, 2021).

A maioria das tarefas de NLP depende de abordagens baseadas em regras projetadas por especialistas em linguagem. A estrutura baseada em regras garante precisão no tocante a uma regra definida, mas possui limitações, já que a inferência é difícil e o custo é muito alto. Com o desenvolvimento contínuo de modelos de DL, vários modelos de aprendizagem também têm sido usados no campo do NLP. A extração e a classificação de relações semânticas estão entre as tarefas de NLP que utilizam modelos de DL, os quais, por serem modelos de classificação, podem garantir dados de resposta corretos e relativamente claros (Jang; Yoon, 2021). As tentativas de aplicar processamento de linguagem natural e aprendizado profundo às tarefas de inteligência tecnológica também progrediram, melhorando o desempenho dos métodos de processamento de texto, como classificação de documentos e recuperação de informações (Kim; Park; Yoon, 2020; Krestel *et al.*, 2021).

No entanto, olhando-se mais adiante, pode-se antecipar que os modelos de DL se tornarão a norma em linguística computacional, com o pré-treinamento e a transferência de aprendizagem desempenhando papéis altamente impactantes. Enquanto arquiteturas mais versáteis e gerais estão obviamente se tornando cada vez mais uma realidade, entender os conceitos abstratos manipulados por tais técnicas é importante para saber como construir e treinar modelos melhores (Babić; Martinčić-Ipšić; Meštrović, 2020).

2.3 REPRESENTAÇÃO DO CONHECIMENTO

A Representação do Conhecimento (do inglês *Knowledge Representation* - KR*)* é um campo da IA que além de tratar de linguagens e técnicas, também se preocupa com a construção de grandes bases de conhecimento (Hinkelmann; Laux, 1993). Segundo Brachman e Levesque (2004), a representação do conhecimento está interessada em mostrar como o conhecimento pode ser representado simbolicamente e manipulado de forma automatizada por softwares capazes de efetuar raciocínio.

Para que o conhecimento seja processado por computador, ele deve ser descrito em uma representação formal por meio de uma linguagem de representação do conhecimento. As linguagens buscam capturar e estruturar conceitos de determinado domínio, ao mesmo tempo que retêm a sua representatividade semântica (Abel; Rama Fiorini, 2013).

Segundo Choi e Kim (2021), a representação do conhecimento é a formalização explícita de premissas e de fenômenos do ambiente do problema em uma linguagem compreensível pela máquina. A KR permite que uma entidade determine consequências ao

raciocinar sobre o espaço do problema em vez de agir, incorporando uma formalidade que pode ser usada para projetar e construir facilmente um sistema complexo.

A representação do conhecimento oferece vários paradigmas, os quais fornecem diferentes mecanismos que permitem, por exemplo, inferências que podem ser usadas na avaliação de conteúdo. Entre os paradigmas mais tradicionais, pode-se citar a lógica de descrição, os *frames*, as redes semânticas e as regras de produção e associação. A lógica de descrição fornece meios para modelar os relacionamentos entre entidades em um domínio específico. É muito utilizada na web semântica e nas ontologias, fornecendo fundamentação para linguagens como a OWL (Web Ontology Language), por exemplo (Krötzsch; Simančík; Horrocks, 2012). No paradigma dos *frames*, as entidades do domínio são representadas na forma de conceitos, propriedades, restrições e instâncias, além de incluir conceitos como herança e *daemons*[21] sobre propriedades (Abel; Rama Fiorini, 2013).

A representação semântica diz respeito ao conjunto de descrições e especificações que reproduzem informações sobre os sujeitos do universo do problema, estrutura/hierarquia/relações dos sujeitos e suas propriedades e funções. Já as regras de produção são úteis para projetar e armazenar informações de uma forma generalizada porque são declarativas e procedimentais. O sistema especialista baseado em regras é o sistema baseado em conhecimento mais antigo e mais conhecido devido à sua asserção simples e à sua fácil implementação para raciocinar sobre asserções (Choi; Kim, 2021).

As regras de produção são uma forma tradicional de representação do conhecimento em sistemas especialistas baseados em regras. Choi e Kim (2021) descrevem várias técnicas de ML que seguem paradigmas diferentes, como ANNs, Árvores de Decisão (do inglês *Decision Tree* - DT) e análise de agrupamentos (*clustering)*. Ainda que distintas das regras de produção, algumas dessas técnicas mais recentes também podem extrair regras a partir de modelos treinados, estabelecendo uma conexão com os sistemas baseados em regras. Dessa forma, dependendo da complexidade e das características do problema a ser modelado, abordagens baseadas em regras e em ML podem ser complementares e integradas.

Já a extração de regras de associação e os padrões fazem o uso de algoritmos de ML para extrair informações latentes de grandes volumes de dados, visando acelerar o processo de construção do conhecimento para a resolução de problemas específicos. Entre as possibilidades, consta a técnica de regras de associação, as quais implementam algoritmos voltados à identificação de regras fortes que governam implicitamente o espaço do problema. É uma

[21] *Daemons* são procedimentos que monitoram com frequência o banco de dados e que disparam quando certas condições são satisfeitas (Brachman; Levesque, 2004).

aprendizagem não supervisionada porque não assume qualquer estrutura; os dados brutos não são anotados ou rotulados. Em muitos casos, portanto, os algoritmos de regra de associação definem o suporte mínimo e os limites de confiança para evitar a captura de regras fortes, mas estatisticamente menos significativas (Choi; Kim, 2021).

As árvores de decisão utilizam uma estrutura de árvore semelhante a um fluxograma para separar um conjunto de dados em várias classes predefinidas, fornecendo assim a descrição, a categorização e a generalização de determinados conjuntos de dados. A partir da representação gerada, permitem a produção de regras que descrevem o espaço de busca. Como um modelo lógico, a árvore de decisão mostra como o valor de uma variável de destino pode ser previsto usando os valores de um conjunto de variáveis preditoras (Yu *et al.*, 2010).

Por fim, cita-se o grande interesse em pesquisas baseadas em grafos de conhecimento com foco na aprendizagem de representação do conhecimento (do inglês *Knowledge Representation Learning* - KRL) ou incorporação de grafos de conhecimento (do inglês *Knowledge Graphs Embeddings* - KGE) com o intuito de mapear entidades e relações em vetores de baixa dimensão enquanto captura seus significados semânticos (Lin *et al.*, 2018; Wang *et al.*, 2017). Um KRL concentra-se principalmente no processo de aprendizagem de *embeddings* de grafos de conhecimento, mantendo as semelhanças semânticas. Outras estruturas que mapeiam a semântica de determinada unidade de informação também têm sido propostas e atualmente promovem certa atenção. Entre elas encontram-se os *embeddings*.

Os conceitos de grafo de conhecimento e de *embeddings* serão assuntos das próximas seções devido ao papel que exercem na proposição desta tese.

2.3.1 Grafo de conhecimento

Em 2012 a Google® incorporou em seu sistema de busca o conceito de grafo do conhecimento (do inglês *Knowledge Graph* - KG) visando melhorar os resultados da sua ferramenta de busca com informações de pesquisa semântica (Singhal, 2012). A ferramenta fornece informações estruturadas e detalhadas sobre o tema, além de uma lista de *links* para outros sites. O objetivo é prover aos usuários as informações necessárias para responder às dúvidas estabelecidas por meio de uma consulta sem ser preciso navegar em outros sites. Segundo a Google®, essa informação é proveniente de várias fontes, incluindo a Freebase®, a Wikipédia® e a CIA World Factbook®, bem como das próprias fontes internas da Google®.

Na literatura há uma vasta lista de definições sobre KG. Ehrlinger e Wöß (2016) trazem uma lista com diversas alternativas e definem um KG como uma descrição estruturada que fornece uma compreensão compartilhada de um determinado domínio.

Um KG é uma representação abstrata que não permite que uma máquina raciocine diretamente. No entanto, quando envolve semântica formal, pode ser utilizada como uma base de conhecimento para interpretação e inferência sobre fatos (Ji *et al.*, 2020).

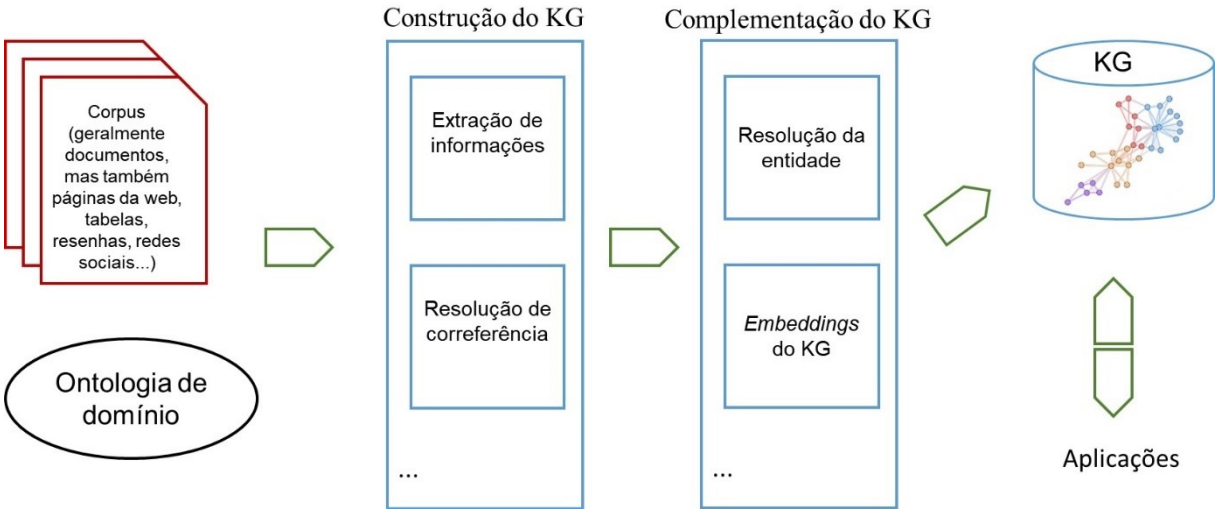
Segundo Deng, Huang e Zhu (2019), os grafos de conhecimento em geral são personalizados para aplicações específicas, sendo recentemente utilizados para apoiar aplicações intensivas em conhecimento, tais como recuperação de informação, resposta automática a perguntas, recomendação personalizada e previsão de tendências de tecnologia. Possuem a capacidade de extrair, organizar e gerenciar efetivamente o conhecimento obtido a partir de dados em larga escala para melhorar a qualidade dos serviços de informação e fornecer aos usuários serviços mais inteligentes (Chen; Jia; Xiang, 2020).

Segundo Ji *et al.* (2020), os KGs permitem uma representação estruturada de fatos, consistindo em entidades, relacionamentos e descrições semânticas. Entidades podem ser objetos do mundo real ou conceitos abstratos, enquanto relacionamentos representam a relação entre entidades. Já descrições semânticas de entidades e seus relacionamentos contêm tipos e propriedades com significado bem definido.

2.3.1.1 Construção de grafo de conhecimento

A construção de um grafo do conhecimento requer algumas metodologias. Kejriwal (2020) apresenta um fluxo simplificado para a construção de um KG de domínio, sendo as etapas principais apresentadas na Figura 8.

Figura 8 – Fluxo simplificado de construção de um KG



Fonte: adaptado de Kejriwal (2020)

As etapas iniciais incluem a aquisição e a limpeza dos dados brutos, assim como o projeto da ontologia de domínio, o qual contém os tipos de entidades de interesse, os relacionamentos que conectam essas entidades e atributos associados às entidades (Kejriwal, 2020).

Na etapa de construção do KG são realizadas a extração de informações (do inglês *Information Extraction* - IE) e a resolução de correferências. A extração de informações envolve extrair as entidades e os relacionamentos (geralmente de forma semiautomática) dos dados brutos (Nadeau; Sekine, 2007). Por exemplo, dado um conjunto de artigos de biologia, um sistema de EI ideal não apenas seria capaz de extrair entidades como proteínas, genes e outras, mas também relações entre elas. É importante notar que o objetivo não é extrair todas as entidades e relações, mas apenas aquelas consideradas relevantes para o domínio (podendo ou não estar especificado na ontologia).

Ainda na construção do KG ocorre a resolução de correferência, que é o problema de resolver "pronomes" e outras palavras e frases para suas menções canônicas em uma tentativa de evitar duplicações desnecessárias e obter dados de maior qualidade.

Já a etapa de complementação de KG é composta pela resolução de entidades e pelos *embeddings* de KG. A resolução de entidades é entendida como um problema que visa agrupar automaticamente entidades que se acredita serem a mesma entidade subjacente (Getoor; Machanavajjhala, 2012; Kejriwal; Miranker, 2015). Após 50 anos de pesquisa, as soluções recentes para o problema têm sido particularmente encorajadoras e se mostraram úteis em arquiteturas do mundo real.

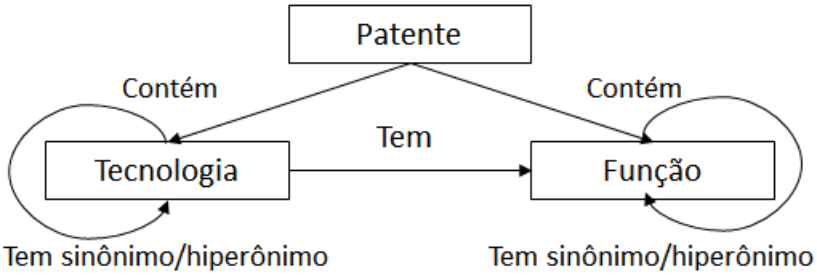
Um KG também pode conter ruído por meio de relações e entidades extraídas incorretamente. Nesse sentido, correções são necessárias, bem como a remoção de *links* errôneos. Para isso, utilizam-se algoritmos de "identificação de grafo de conhecimento". Esses algoritmos contam com uma variedade de metodologias (Pujara *et al.*, 2013), mas, recentemente, técnicas de DL, como geração de *embeddings* de KG, provaram ser soluções eficazes e de rápido aprimoramento (Wang *et al.*, 2017).

Por fim, com o KG construído, este é normalmente armazenado em uma infraestrutura responsiva e especializada, como um banco de dados orientado a grafos. Em seguida, deve ser exposto a aplicações como mecanismos de pesquisa e interfaces de resposta a perguntas (como, por exemplo, *chatbot*) (Kejriwal, 2020).

Direcionando o estudo para patentes, Deng, Huang e Zhu (2019) afirmam que o procedimento de construção de um KG voltado ao cenário de patentes compreende três etapas principais. A primeira etapa procura definir o esquema do KG da patente que especifica entidades e relacionamentos importantes para a compreensão dos documentos técnicos. A segunda etapa objetiva criar um conjunto de dados rotulados, anotando manualmente os documentos de patentes com tipos de entidade e relacionamentos. Essa etapa não é necessária se já houver dados rotulados suficientes. Por fim, a última etapa visa treinar o modelo de extração de conhecimento com os dados rotulados e usar o modelo treinado para extrair possíveis instâncias de entidade e relacionamentos de novos documentos de patentes. As entidades e relacionamentos extraídos são então armazenados no formato de triplas sujeito-predicado-objeto, que constituem o grafo de conhecimento da patente, ou mesmo diretamente em uma estrutura de grafo provida por banco de dados voltados para tal.

No artigo de Deng, Huang e Zhu (2019), define-se um esquema de KG de patentes de utilidade (patentes de design, de plantas e de utilidades), como apresentado na Figura 9.

Figura 9 – Esquema de grafo do conhecimento



Fonte: adaptado de Deng, Huang e Zhu (2019)

Uma patente contém uma ou mais tecnologias, que podem ser novas soluções, sistemas, produtos, entre outras. Cada tecnologia possui uma ou mais funções, que podem ser seus atributos, objetivos, capacidades e assim por diante. A partir de um documento de patente, são extraídas as principais tecnologias, funções, a relação de dependência entre funções e tecnologias bem como suas relações de sinônimos ou hiperônimos (Deng; Huang; Zhu, 2019).

Atualmente, a pesquisa em KG concentra-se principalmente em três aspectos: 1) representação de conhecimento, 2) construção e 3) aplicação de grafo de conhecimento, que integra computação cognitiva, representação de conhecimento e raciocínio, recuperação e extração de informação, processamento de linguagem natural, mineração de dados e outras áreas e tecnologias (Chen *et al.*, 2020).

2.3.2 Embeddings

A representação ou incorporação (*embeddings*) de vetores distribuídos vem ganhando destaque em áreas como NLP e DL. Isso ocorre porque representações mais tradicionais, baseadas na frequência da palavra, principalmente aplicadas para classificação de texto, como *Bag-of-Word* (BOW), *Term Frequency-Inverse Document Frequency* (TF-IDF) ou o modelo de *N-Grams,* ignoram a ordem das palavras e suas semelhanças sintáticas e semânticas (Kalyan; Sangeetha, 2020; Li *et al.*, 2018).

Para lidar com a limitação dos recursos de frequência de palavras, um vetor de palavras foi proposto e demonstrado por Mikolov *et al.* (2013), no qual se realizou a captura de relações sintáticas e semânticas. Com isso, tornou-se possível melhorar a qualidade dos vetores e acelerar o treinamento, como a subamostragem de palavras frequentes. De modo geral, os *embeddings* mapeiam as palavras em representações vetoriais densas, bem como capturam informações sintáticas e semânticas minimizando questões como o problema da dimensionalidade e a falta de informações nas representações (Kalyan; Sangeetha, 2020).

Os autores Kalyan e Sangeetha (2020) classificaram *embeddings* em duas categorias, dependendo do tipo de dados que estes mapeiam: 1) *embeddings* de texto ou 2) *embeddings* de conceito. Visto que este trabalho se preocupa com a classificação de patente como suporte para a recomendação de subclasses, tem-se ênfase nos *embeddings* de texto, que, dependendo da granularidade, podem ser divididos em cinco tipos: 1) *character embeddings*; 2) *word embeddings*; 3) *phrase embeddings*; 4) *sentense embeddings;* e 5) *document embeddings*. Além dos tipos citados, Krestel *et al.* (2021) destacam também o *graph embeddings* e o *contextual word embeddings*.

Os modelos de *character embeddings* consideram os caracteres como uma unidade atômica e os mapeiam para um vetor denso de comprimento fixo. Os vetores ao nível do caractere codificam informações morfológicas como prefixo e sufixo, bem como informações ortográficas. Eles oferecem vetores de qualidade em tarefas que necessitam lidar com o conceito de OOV (*Out of Vocabulary*), bem como com palavras raras (Kalyan; Sangeetha, 2020).

Os *word embeddings* são vetores densos aprendidos a partir de grandes coleções de texto que observam o contexto de cada palavra (Camacho-Collados; Pilehvar, 2018). Um método de incorporação muito difundido é o *word2vec* (Mikolov *et al.*, 2013b), que aprende representações vetoriais de palavras de maneira não supervisionada. Ou seja, por meio do treinamento em uma grande coleção de documentos, palavras com significado semelhante são atribuídas a vetores que permitem identificar essa semelhança. Os modelos mais conhecidos e utilizados de *word2vec* são o Continuous Bag of Word (CBOW) e o Skip-gram.

Quando se trata de representar frases, sentenças, parágrafos ou documentos inteiros, vários métodos foram propostos. A abordagem mais simples utiliza a média dos vetores de incorporação de palavras de um texto, ignorando a ordem em que as palavras aparecem. Dado que o significado de uma frase é mais do que a soma do significado de suas palavras, essa abordagem simples não pode capturar diferenças semânticas sutis ao nível de frase (Krestel *et al.*, 2021).

Nesse sentido, um modelo bastante utilizado é o *paragraph2vec*, mais conhecido como *doc2vec*, sendo uma extensão do *word2vec*. Trata-se de um modelo não supervisionado que mapeia texto de comprimento variável como frases, parágrafos e documentos para representações vetoriais densas. *Doc2vec* aprende representações de vetores densos para texto de comprimento variável e palavras no *corpus*, oferecendo dois modelos: 1) *Distributed Bag of Words* (DBOW) e 2) *Distributed Memory* (Kalyan; Sangeetha, 2020).

Já o *graph embeddings* é uma técnica de ML que mapeia os nós e as arestas de um grafo em um espaço vetorial de baixa dimensão. O objetivo é aprender representações vetoriais significativas para os nós que capturem as propriedades topológicas e relacionais do grafo (Krestel *et al*., 2021). A ideia de representar dados de texto como vetores semanticamente significativos também foi adotada para representar dados de grafo, por exemplo, no algoritmo *node2vec* (Grover; Leskovec, 2016).

Os métodos mais recentes de DL para representar dados textuais são os *contextual graph embeddings,* que criam representações altamente específicas do contexto de cada palavra. A contextualização é frequentemente empregada em DLs que utilizam modelos de linguagem

como BERT (Devlin *et al.*, 2019), ELMo (Peters *et al.*, 2018) e GPT (Brown *et al.*, 2020), discutidos na seção 2.4.1.4.

Por fim, o *sentence embedding* refere-se ao processo de representar uma frase ou sentença em um espaço vetorial. Os *embeddings* de sentenças são usados em várias aplicações, com destaque para classificação de texto, recuperação de informações, resumo automático e tradução automática. São úteis para comparar a semelhança entre frases e entender o contexto (Reimers; Gurevych, 2019).

2.3.2.1 Approximate Nearest Neighbor (ApNN[22])

A representação de textos em espaços vetoriais densos, também conhecidos como *embeddings* textuais, tem sido amplamente estudada nos últimos anos devido à sua capacidade de capturar relações sintáticas e semânticas entre palavras, documentos e outras unidades de informação. Alguns métodos para geração de *embeddings* textuais, apresentados na seção anterior, são utilizados para recuperar palavras e documentos semanticamente similares em um espaço vetorial, o que constitui, essencialmente, uma busca por vizinhos mais próximos (Renga Bashyam; Vadhiyar, 2020).

Segundo Renga Bashyam e Vadhiyar (2020), a técnica *Approximate Nearest Neighbor* (ApNN) tem como objetivo encontrar objetos em um conjunto de dados que sejam mais semelhantes a um objeto de consulta. No entanto, a busca exata de ApNN em conjuntos de dados de alta dimensionalidade pode ser computacionalmente dispendiosa. Nesse sentido, as soluções atuais visam melhorar a escalabilidade e a eficiência da busca de ApNN em grandes conjuntos de dados.

Já Li *et al*. (2020) definem ApNN como uma maneira de encontrar o vizinho mais próximo de um determinado ponto em um conjunto de dados, de forma aproximada em vez de exata. Isso é útil quando o conjunto de dados é muito grande e a busca exata tem custo alto em termos de tempo e de recursos computacionais.

A ApNN é realizada por meio de técnicas de *hashing* como o algoritmo *Locality-Sensitive Hashing* (LSH), métodos baseados em árvores (KD-*trees*, R-*trees* e VP-*trees*) e métodos baseados em grafos (Renga Bashyam; Vadhiyar, 2020). O LSH é uma técnica que mapeia pontos de dados em um espaço de alta dimensão para um espaço de baixa dimensão, de

[22] Será utilizado ApNN para representar *Approximate Nearest Neighbor*, de modo a diferenciar de ANN (do inglês *Artificial Neural Network*).

modo que pontos próximos no espaço de alta dimensão sejam mapeados para pontos próximos no espaço de baixa dimensão (Malkov *et al*., 2014).

Os métodos baseados em árvores aceleram o processo de busca, particionando o espaço hierarquicamente em uma estrutura de árvore e dividindo os pontos ao longo da estrutura. Embora esses métodos reduzam o tempo de busca substancialmente para conjuntos de dados de baixa dimensão, seu desempenho começa a sofrer à medida que o número de dimensões aumenta porque as distâncias entre os pontos não são distintas o suficiente (Kim *et al*., 2023).

Outra possibilidade é por meio do algoritmo Hierarchical Navigable Small World (HNSW), que possibilita a busca aproximada de vizinhos baseada em grafos. A ideia central do HNSW é separar as ligações de acordo com sua escala de distância em diferentes camadas, e então realizar a busca em um grafo de múltiplas camadas. Ele mantém poucas conexões de longo alcance na camada superior, enquanto conexões densas e de curto alcance estão nas camadas inferiores, o que reduz o tempo de comparações e localizações (Kim *et al.*, 2023).

Por fim, a ApNN possui diversas aplicações, como classificação, agrupamento, redução de dimensionalidade, reconhecimento de objetos, recuperação de imagens, busca de similaridade de documentos e sistemas de recomendação. Ademais, a escolha da técnica de ApNN e de métricas de avaliação adequadas depende de aplicação específica (Kim *et al*., 2023; Li *et al.*, 2020; Renga Bashyam; Vadhiyar, 2020).

2.4 APRENDIZADO PROFUNDO

A Inteligência Artificial (do inglês *Artificial Intelligence* - AI), o Aprendizado de Máquina (do inglês *Machine Learning* - ML) e o Aprendizado Profundo (do inglês *Deep Learning* - DL) são áreas e subáreas emergentes às vezes usadas indistintamente para descrever sistemas ou softwares que se comportam de maneira inteligente (Sarker, 2021a), conforme representado na Figura 10. Resumidamente, a AI incorpora o comportamento humano e a inteligência às máquinas ou sistemas (Sarker; Furhad; Nowrozy, 2021), enquanto o ML é um subconjunto da AI que se utiliza de métodos para aprender a partir de dados ou experiência, possibilitando automatizar a construção de modelos analíticos (Sarker, 2021a).

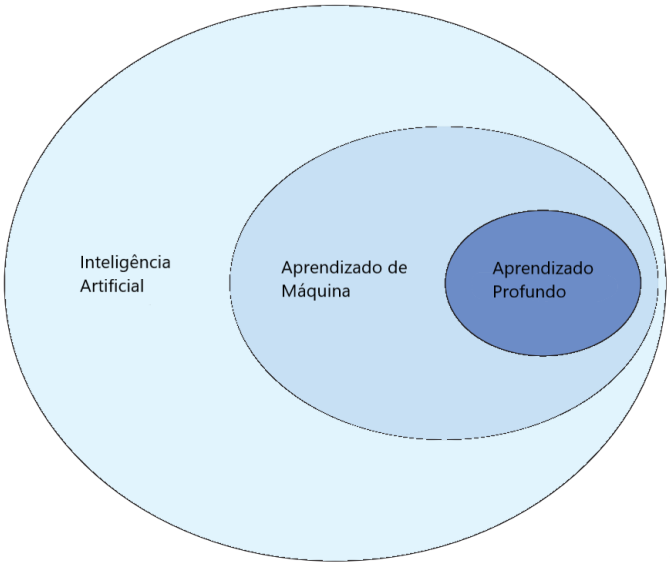
Já a DL também é considerada um subcampo específico do ML e representa métodos de aprendizagem de dados, sendo o processamento realizado por meio de redes neurais multicamadas (Alom *et al.*, 2019; Sarker, 2021b). Segundo Chollet (2017), a DL é uma

abordagem mais recente sobre o aprendizado de objetos a partir de dados e enfatiza o aprendizado de camadas sucessivas, buscando representações cada vez mais significativas.

O termo "*deep*" em *deep learning* diz respeito à quantidade de camadas que contribuem para um modelo de dados, chamada de profundidade do modelo. As camadas são aprendidas automaticamente com a exposição aos dados de treinamento (Chollet, 2017; Russell; Norvig; Chang, 2021). Essas representações em camadas são (quase sempre) aprendidas por meio de modelos de redes neurais, estruturados em camadas empilhadas umas sobre as outras (Chollet, 2017).

Segundo Alom *et al.* (2019), os resultados experimentais mostram que a DL apresenta desempenho superior quando comparada com os métodos tradicionais de ML em áreas como processamento de imagem, visão computacional, reconhecimento de fala, tradução automática, arte, imagens médicas, processamento de informações médicas, robótica e controle, bioinformática, processamento de linguagem natural, cibersegurança e muitas outras.

Figura 10 – Relação entre Inteligência Artificial, Aprendizado de Máquina e Aprendizado Profundo



Fonte: Teofili (2019, p. 5, tradução nossa)

Além das áreas citadas, a DL vem sendo utilizada para analisar dados de documentos de patentes. As coleções de documentos de patentes fornecem uma grande fonte de conhecimento, e o volume de dados cresce rapidamente, tornando-se um desafio recuperar e analisar informações dessa fonte de maneira eficaz. Com base em técnicas de DL, abordagens vêm sendo desenvolvidas no campo da análise de patentes e têm como objetivo a redução de

custos e a automação de tarefas que apenas especialistas de domínio eram capazes de realizar (Krestel *et al.*, 2021).

No domínio das patentes, classificá-las é uma tarefa que requer automação para auxiliar os especialistas na organização e na recuperação de informações. As patentes são categorizadas de acordo com esquemas-padrão de classificação, como a taxonomia IPC, que possui uma hierarquia abrangendo todos os campos tecnológicos. Uma patente pode ser categorizada em uma ou mais dessas áreas e, portanto, a classificação de patentes é um problema de classificação com vários rótulos. Nesse sentido, capturar os recursos de cada categoria e aprender os espaços de incorporação do texto tornam-se tarefas desafiadoras (Roudsari *et al.*, 2021).

Diferentes métodos e técnicas têm sido propostos com base em abordagens distintas de aprendizagem, incluindo aprendizagem supervisionada, semissupervisionada, não supervisionada e por reforço (do inglês *Reinforcement Learning* - RL) (Alom *et al.*, 2019).

A aprendizagem supervisionada profunda (do inglês *Deep Supervised Learning* - DSL) é um tipo de treinamento em que se especifica a saída desejada correspondente a cada entrada (Teofili, 2018). O modelo aprende observando os pares de entrada-saída de exemplos fornecidos por um "professor" (Russell; Norvig; Chang, 2021).

Segundo Alom *et al.* (2019), após o treinamento bem-sucedido, o modelo será capaz de fornecer respostas às perguntas, ou seja, proverá uma saída a partir de uma entrada. Existem diferentes arquiteturas de aprendizagem supervisionada para a DL, incluindo as seguintes: Redes Neurais Convolucionais (do inglês *Convolutional Neural Network* - CNN); Redes Neurais Recorrentes (do inglês *Recurrent Neural Network* - RNN); Memória de Longo Prazo (do inglês *Long Short-Term Memory* - LSTM), sendo esta um tipo de RNN; e *Multilayer Perceptron* (MLP).

Já a aprendizagem semissupervisionada profunda (do inglês *Deep Semi-Supervised Learning* - DSSL) foi projetada para dados parcialmente rotulados. As redes RNN e LSTM também são usadas para aprendizagem semissupervisionada (Mishra *et al.*, 2020).

Na aprendizagem não supervisionada profunda (do inglês *Deep Unsupervised Learning* - DUL), o modelo aprende padrões na entrada sem qualquer *feedback* explícito. O agrupamento de dados (*clustering*) é a tarefa de aprendizagem não supervisionada mais comum que detecta grupos potencialmente úteis de exemplos a partir de um conjunto de entrada (Russell; Norvig; Chang, 2021). Frequentemente agrupamentos, redução de dimensionalidade e técnicas generativas são considerados conceitos envoltos na aprendizagem não supervisionada. Existem métodos e técnicas de aprendizado profundo mais adequados para a

realização de agrupamento e redução de dimensionalidade não linear, como, por exemplo, Auto-Encoders (AE), Restricted Boltzmann Machines (RBM) e Generative Adversarial Networks (GAN). Além disso, RNNs, como LSTM e RL, também são usados para aprendizagem não supervisionada em muitos domínios de aplicação (Alom *et al.*, 2019).

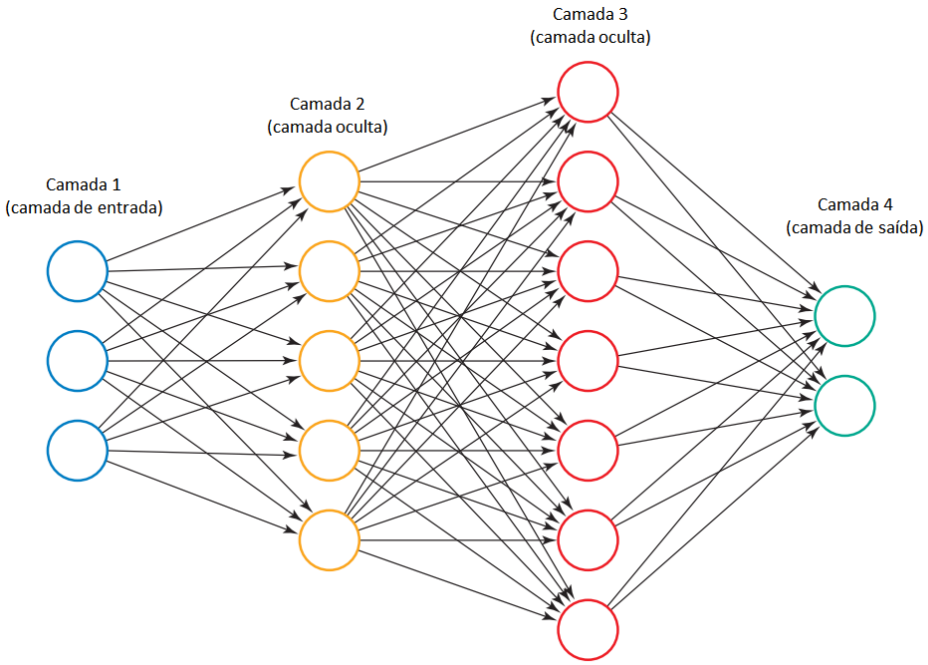
Por fim, o RL é um método que recebe informações sobre seu ambiente e aprende a escolher ações que vão maximizar alguma recompensa (Chollet, 2017). Um agente interage com um ambiente, recebe recompensas ou penalidades em resposta às suas ações e aprende a se comportar, de modo a maximizar sua recompensa total ao longo do tempo (Russell; Norvig; Chang, 2021).

Como se pode perceber, o aprendizado profundo tem sua origem nas redes neurais, e nos últimos anos vários métodos foram desenvolvidos para acompanhar a evolução, principalmente no NLP e no CV. Nas próximas seções serão apresentadas algumas arquiteturas de aprendizado profundo, tais como CNN, RNN, LSTM e MLP.

2.4.1 Arquiteturas de redes neurais profundas

As redes neurais artificiais profundas (do inglês *Deep Neural Network* - DNN) representam um paradigma computacional originalmente inspirado pela maneira como o cérebro é organizado em grafos de neurônios (embora o cérebro seja muito mais complexo do que uma rede neural artificial) (Teofili, 2018). Uma DNN geralmente consiste em uma camada de entrada, uma ou mais camadas ocultas e uma camada de saída em que cada camada é composta por vários neurônios (Zhang *et al.*, 2021). A profundidade da rede ocorre pela quantidade de camadas ocultas. A Figura 11 apresenta uma visão geral de DNN, com duas camadas ocultas:

Figura 11 – DDN com duas camadas ocultas



Fonte: adaptado de Teofili (2018, p. 6)

Segundo Teofili (2018), as redes neurais profundas são mais adequadas para:

- fornecer uma representação de dados textuais que captura a semântica de palavras e documentos, permitindo que uma máquina entenda quais palavras e documentos são semanticamente semelhantes;
- gerar texto que seja significativo em um determinado contexto, por exemplo, para criação de *chatbots*;
- fornecer representações de imagens que não pertencem aos pixels, mas sim aos seus objetos de composição. Isso habilita a construção de sistemas eficientes de reconhecimento facial e imagens em geral; e
- executar a tradução automática de textos com eficiência.

Segundo Krestel *et al.* (2021), os termos *aprendizado profundo* e *redes neurais* profundas geralmente se referem às arquiteturas CNN e RNN, respectivamente. A diferença entre essas duas arquiteturas é que elas são adaptadas para reconhecer diferentes tipos de padrões nos dados. Além das redes CNN e RNN, serão abordadas também a LSTM e a MLP.

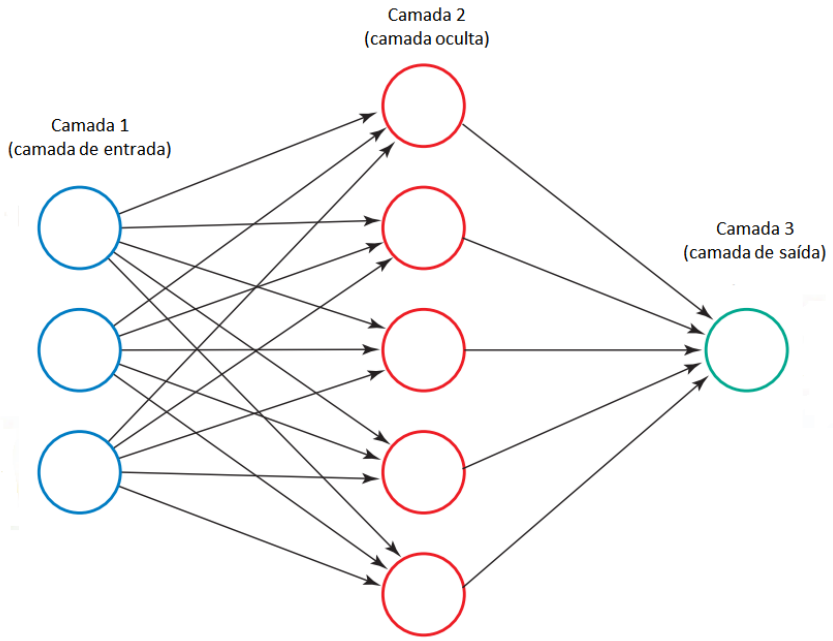
2.4.1.1 Multilayer Perceptron (MLP)

A rede neural *Multilayer Perceptron* (MLP) é uma arquitetura de ANN que possui uma abordagem supervisionada, sendo utilizada para resolver problemas de classificação e previsão (Sarker, 2021a). Para Russell, Norvig e Chang (2021), a MLP é uma rede neural na qual as conexões possuem uma única direção, formando um grafo acíclico sem *loops*.

A estrutura típica de um MLP possui três camadas – entrada, oculta e saída –, conforme apresentado na Figura 12. A camada de entrada consiste no número de variáveis de entrada de um conjunto de dados. Uma ou mais camadas ocultas aceitam o conteúdo da camada de entrada. Por fim, a camada de saída recebe a saída da camada oculta e produz o rótulo de classe ou previsão (Kumar; Ravi, 2016).

A MLP utiliza o algoritmo de retropropagação (*backpropagation*), onde a camada oculta tem o intuito de atualizar os pesos e ativar funções (não lineares) que auxiliam a capturar a não linearidade dos dados (Kumar; Ravi, 2016). Os principais casos de uso do MLP são classificação, reconhecimento de padrões, previsão e aproximação (Abirami; Chitra, 2020).

Figura 12 – Rede neural MLP



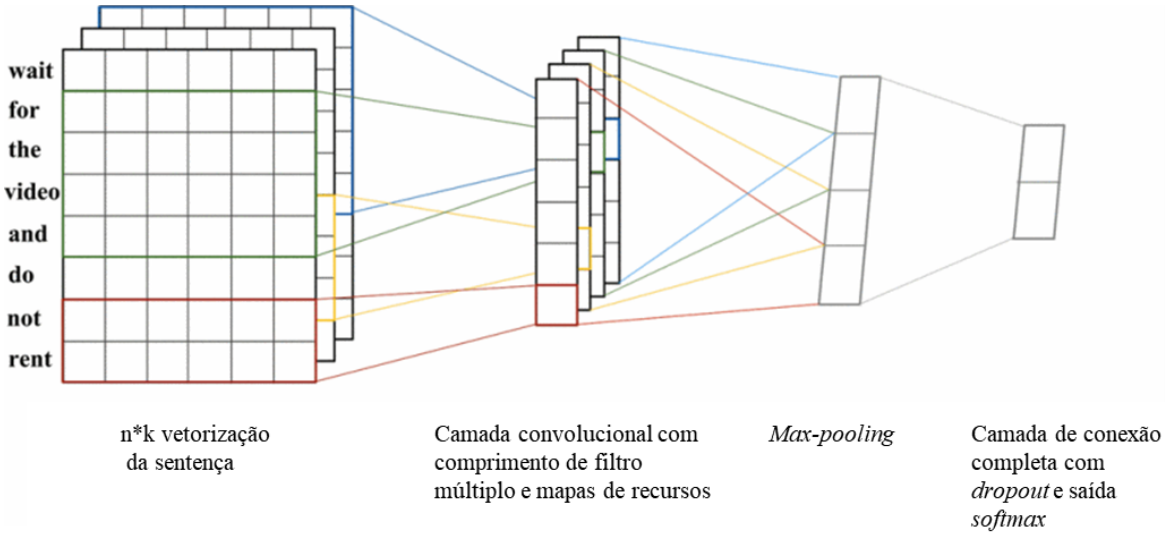
Fonte: adaptado de Teofili (2019)

2.4.1.2 Convolutional Neural Network (CNN)

As Redes Neurais Convolucionais (do inglês *Convolutional Neural Network* - CNN) têm sido extremamente bem-sucedidas em aplicações práticas, principalmente no que se refere à Visão Computacional (do inglês *Computer Vision* - CV). O nome "rede neural convolucional" indica que a rede emprega uma operação matemática chamada de convolução, entendida como um tipo especializado de operação linear. As redes convolucionais são redes neurais que usam convolução no lugar da multiplicação geral da matriz em pelo menos uma de suas camadas (Goodfellow; Bengio; Courville, 2016).

A CNN é amplamente utilizada em visão computacional, classificação de imagens, reconhecimento de fala e NLP em razão da capacidade de capturar correlações locais de estruturas espaciais ou temporais (Hu *et al*., 2018). Wang, Y. *et al.* (2018) utilizaram uma rede baseada em CNN para classificação de sentenças, visto que esse tipo de rede possui um bom desempenho na classificação de texto. Como pode ser observado na Figura 13, de modo geral a CNN consiste em quatro camadas: 1) incorporação de palavras (camada de entrada); 2) convolução; 3) camada de agrupamento (*max-pooling*); e 4) conexão completa (camada de saída),

Figura 13 – CNN para classificação de sentenças



Fonte: adaptado de Wang, Y. *et al.* (2018)

A rede CNN, apesar de sua utilização primária em CV, é também empregada para classificar textos, inclusive os de patentes. A primeira camada do modelo converte as palavras em vetores de baixa dimensão. As próximas duas camadas extraem recursos avançados desses

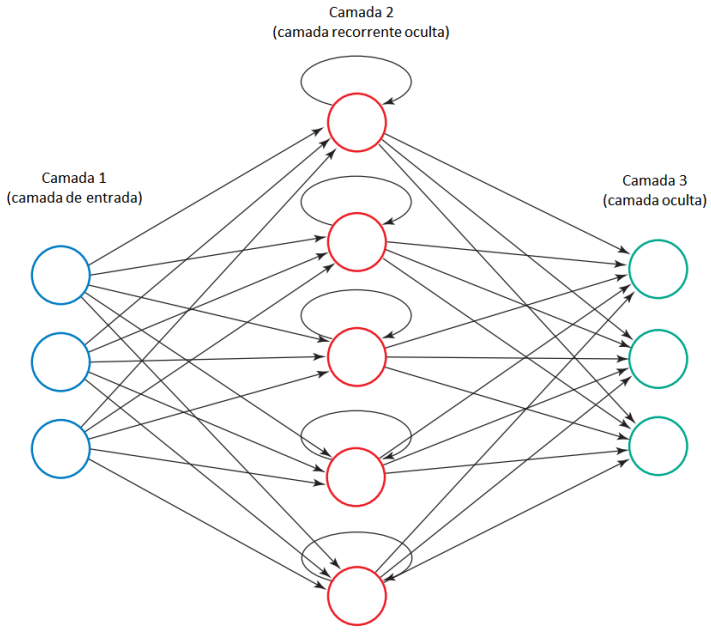
vetores. A camada de conexão completa adiciona regularização de *dropout* e classifica o resultado usando uma camada *softmax*. Cada sentença do texto da patente é representada como uma matriz $n \times k$, onde n é o número total de palavras em uma sentença e k é a dimensão do vetor original da palavra. O modelo usa filtros com diferentes comprimentos e um número razoável de filtros para obter vetores de recursos de k dimensões (Wang, Y. *et al*., 2018).

2.4.1.3 Recurrent Neural Network (RNN)

Uma Rede Neural Recorrente (do inglês *Recurrent Neural Network* - RNN) é uma arquitetura que possibilita o aprendizado profundo para processar dados sequenciais e compartilhar seus parâmetros de modelo em toda a sequência. Essencialmente, as RNNs foram elaboradas para trabalhar com sequências de dados (séries temporais, sequências de texto). Desse modo, podem memorizar as partes anteriores de uma sequência para a classificação das partes posteriores (Krestel *et al.*, 2021).

As RNNs utilizam vetores para a entrada (Camada 1) e a saída (Camada 3), enquanto a camada recorrente oculta (Camada 2) combina o sinal de entrada e um sinal armazenado internamente (*loop* da Camada 2), que desempenha o papel de memória. Essas camadas estão representadas na Figura 14.

Figura 14 – RNN para aprendizagem de sequência



Fonte: adaptado de Teofili (2018, p. 130)

A camada central é responsável por transformar a entrada em saída. O neurônio recorrente (camada 2) combina o sinal do neurônio de entrada (camada 1) com um sinal armazenado internamente (representado pela seta de *loop*), desempenhando assim o papel de memória por meio dessa combinação. Os neurônios processam a entrada, transformando-a em uma saída, dado seu estado interno (os pesos da camada oculta e a função de ativação), atualizando seu estado como uma nova função da nova entrada e seu estado atual. Dessa forma, o neurônio aprende a relacionar as entradas subsequentes, ou seja, durante o treinamento a rede aprenderá quais as palavras mais significativas que têm a probabilidade de aparecer próximas (Teofili, 2018).

Já a camada de saída da RNN utilizada para geração de texto produz um vetor contendo um número real (entre 0 e 1) para cada unidade de saída possível. Esse número representa a probabilidade de determinada unidade ser enviada pela rede, sendo a função *softmax* responsável pela geração das probabilidades. A camada central (Camada 2) é a camada recorrente, responsável por lembrar as sequências vistas anteriormente (Teofili, 2018).

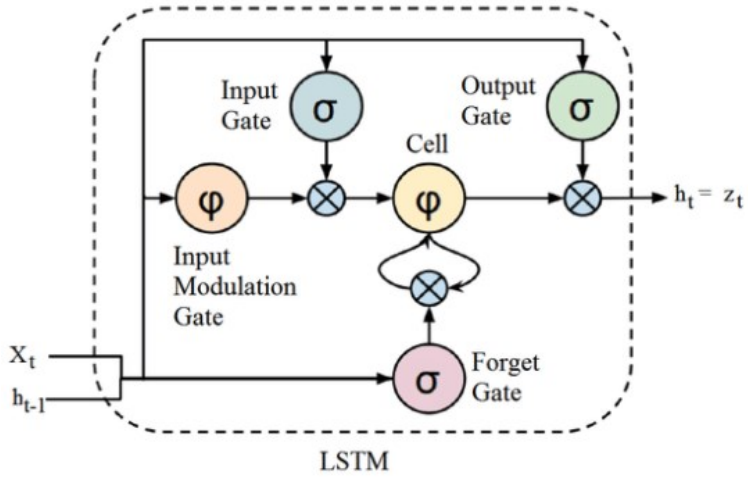
A RNN possui duas extensões complexas, normalmente baseadas em células *long short-term memory network* (LSTM) ou *gated recurrent units* (GRUs). Neste trabalho, a ênfase será na rede LSTM por ser a mais utilizada em tarefas que lidam com textos.

2.4.1.3.1 Long Short-Term Memory Network (LSTM)

Redes neurais recorrentes, como Memória de Longo e Curto Prazo (do inglês *Long Short-Term Memory* - LSTM), configuram uma rede neural profunda típica utilizada para modelar relacionamentos entre sequências de dados em vez de apenas entradas fixas (Goodfellow; Bengio; Courville, 2016). Uma rede LSTM típica é composta por blocos de memória chamados de células, que servem para armazenar o histórico de informações. A atualização e o uso do histórico de informações são controlados por três portas: 1) a porta de entrada; 2) a porta de esquecimento; 3) e a porta de saída, como apresenta a Figura 15 (Data Science Academy, 2022).

Os dados podem ser adicionados ou removidos do estado da célula através de portas sigmoides. Uma porta é semelhante a uma camada ou a uma série de operações de matriz e contém diferentes pesos individuais. As LSTMs são projetadas objetivando evitar o problema de dependência de longo prazo porque usam portas para controlar o processo de memorização (Wang, D. *et al*., 2018).

Figura 15 – Rede neural LSTM



Fonte: Data Science Academy (2022)

Segundo Hong e Wang (2021), a arquitetura de rede LSTM é amplamente usada para classificação de texto, reconhecimento de entidades, correspondência de texto posicional, resumo de texto, análise de similaridade de texto e análise de sentimentos.

2.4.1.4 Transformers

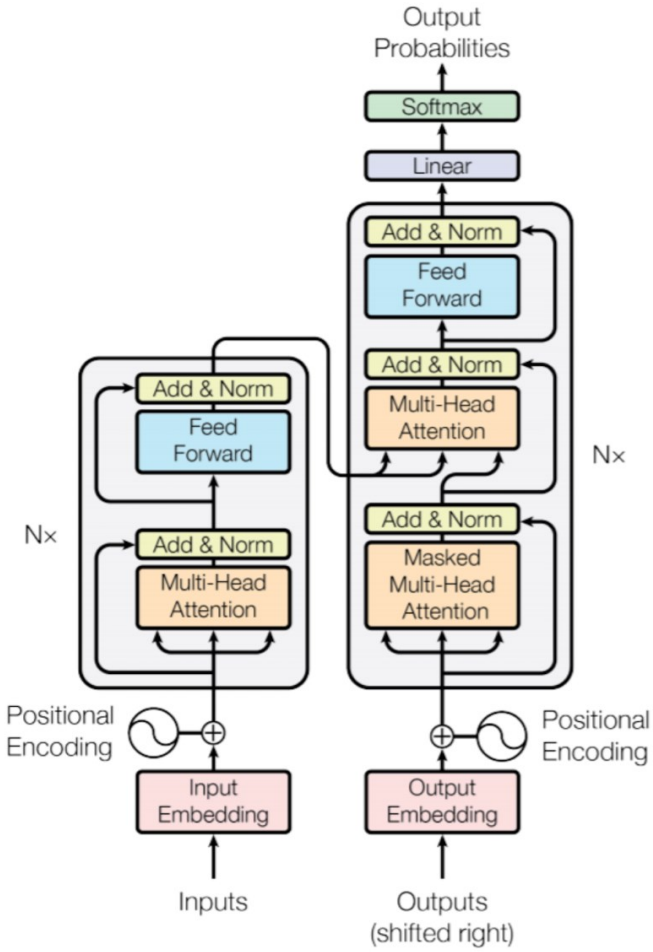
Em 2017, uma equipe de pesquisadores da Google® propôs uma arquitetura de rede neural baseada no conceito de transformador (do inglês *Transformer Neural Network* – TNN, ou simplesmente *transformer*). Consiste em um modelo de DL que utiliza uma arquitetura codificador-decodificador baseada no mecanismo de autoatenção (Vaswani *et al*., 2017). O mecanismo de autoatenção tem como objetivo extrair a relação/dependência entre as palavras da sentença de entrada, auxiliando na obtenção da compreensão global do contexto (Xu; He; Li, 2020).

A rede neural RNN, por exemplo, faz a leitura sequencial da entrada, enquanto o *transformer* realiza a leitura da sequência de entrada de uma só vez. Essa característica permite que o modelo aprenda considerando tanto o contexto anterior quanto posterior de uma palavra (Xu; He; Li, 2020).

A Figura 16 representa a arquitetura do *transformer* apresentada por Vaswani *et al*. (2017), a qual se utiliza de camadas de autoatenção empilhadas e camadas totalmente conectadas por pontos para o codificador e decodificador. Segundo os autores o codificador é uma pilha de 6 camadas idênticas, e cada camada consiste em uma subcamada de autoatenção e uma subcamada de rede *feed-forward*. O codificador recebe uma sequência de vetores

(empacotados em uma matriz) como entrada, processa os vetores com a subcamada de autoatenção e os envia para a camada de rede *feed-forward*. Esta, por sua vez, encaminha os vetores como saída para a próxima camada do codificador.

Figura 16 – Arquitetura do *transformer*



Fonte: Vaswani *et al*. (2017)

Já ao decodificador, que possui estrutura idêntica ao codificador, é adicionada uma terceira subcamada entre as subcamadas de autoatenção e a camada de rede *feed-forward*. Essa terceira camada realiza a atenção, ou seja, o aprendizado dos diferentes tipos de relações de uma sentença, por exemplo, utilizando-se de várias cabeças (entendidas com subcamadas) sobre a saída da pilha do codificador. De modo semelhante ao codificador, empregam-se conexões residuais em torno de cada uma das subcamadas, seguidas pela normalização dessas subcamadas (Vaswani *et al*., 2017).

A arquitetura *transformer* é muito utilizada para resolver problemas de NLP, o que contribuiu para a evolução de modelos pré-treinados em grandes conjuntos de dados não

rotulados, tais como Bidirectional Encoder Representations from Transformers (BERT) e Generative Pre-Trained Transformer (GPT) (Lauriola; Lavelli; Aiolli, 2022).

2.4.1.4.1 LLM

Os modelos BERT e GPT vistos anteriormente são exemplos de modelos de linguagem pré-treinados (*Pre-Trained Models* - PTMs) baseados na arquitetura *transformer*, sendo considerados importantes para a evolução dos modelos de linguagem. Ambos os modelos são exemplos de Modelo de Linguagem de Larga Escala (do inglês *Large Language Models* - LLM), pois contêm um grande número de parâmetros treinados para o objetivo de atender diferentes tarefas de processamento de texto.

A abordagem LLM consiste no treinamento de modelos de linguagem pré-treinados em grandes conjuntos de dados não rotulados, visando adquirir representações de linguagem altamente generalizáveis que podem ser adaptadas para tarefas específicas de NLP (Zhao *et al*., 2023).

Os LLMs possuem a capacidade de aprender e prever a próxima palavra ou caractere em uma sequência de texto, além de serem capazes de gerar texto coerente e executar tarefas de linguagem natural, como tradução automática, respostas a perguntas e geração de texto, bem como geração de *embeddings*. O GPT-3, por exemplo, é um dos maiores modelos de linguagem já criados, com 175 bilhões de parâmetros, sendo capaz de realizar tarefas de linguagem natural com poucas amostras ou instruções simples (Brown *et al*., 2020).

Recentemente a OpenAI® apresentou o modelo GPT-4, adicionando melhorias na capacidade de segurança, comparado às versões anteriores do GPT. Testes abrangendo uma ampla gama de tarefas complexas geradas por humanos mostraram que o GPT-4 consegue resolver esses problemas difíceis de modo significativamente superior ao de seus predecessores (Zhao *et al*., 2023).

Além de BERT e GPT, outros LLM utilizam a arquitetura *transformer*, a saber: T5 (*Text-to-Text Transfer Transformer*) (Raffel *et al*., 2020); RoBERTa (*Robustly Optimized* BERT *Pretraining Approach*) (Liu *et al*., 2019); XLNet (*eXtreme Language understanding Network*) (Yang *et al*., 2019); ALBERT (*A Lite BERT*) (Lan *et al*., 2020); e SBERT (*Sentence*-BERT) (Reimers; Gurevych, 2019).

Segundo Zhao *et al*. (2023), os LLMs apresentam uma série de desafios e limitações, entre os quais se destaca a necessidade de um volume considerável de dados e de recursos computacionais para o treinamento desses LLMs. Além disso, a interpretação dos resultados

gerados pelos LLMs pode ser complexa e apresentar dificuldades. Os LLMs podem também ter dificuldade em lidar com nuances e ambiguidades na linguagem natural e podem gerar respostas que não fazem sentido ou são inadequadas para uma determinada tarefa. O tema é recente, tratado na literatura como “alucinação” em LLMs.

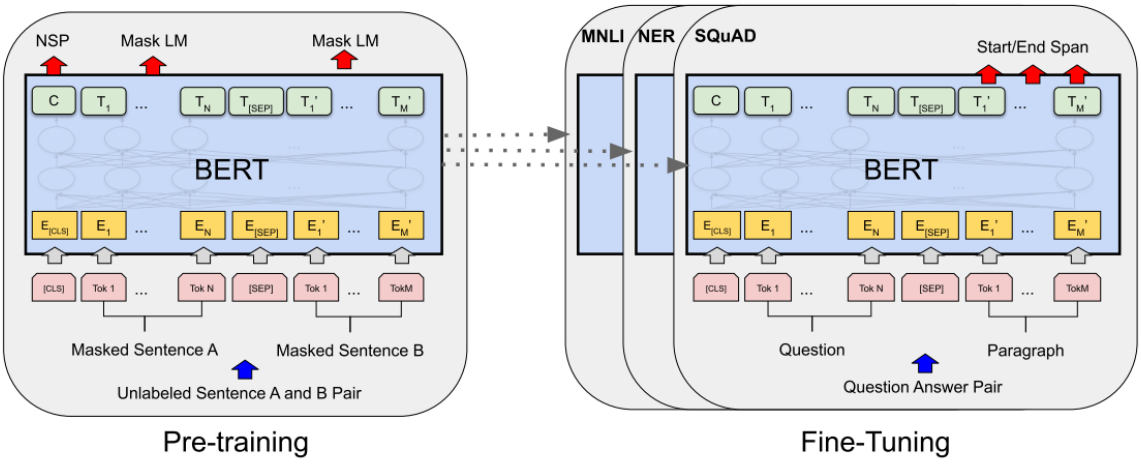
Zhao *et al*. (2023) afirmam que para minimizar as “alucinações” em LLMs algumas estratégias têm sido propostas, entre elas: a) realização de *fine-tuning*, treinando o LLM com dados de alta qualidade alinhados às tarefas desejadas, visando reduzir a discrepância entre os objetivos do modelo e do usuário; b) utilização de *feedback* humano durante o treinamento; c) treinamento de modelos separados para detecção de alucinação, filtrando saídas incorretas. Dessa forma, o foco é alinhar objetivos e capacidades do LLM ao uso desejado, reduzindo a probabilidade de alucinações.

2.4.1.4.2 BERT

Em 2018, o Departamento de Inteligência Artificial da Google® publicou o modelo de aprendizado profundo pré-treinado chamado BERT (Bidirectional Encoder Representations from Transformers). Segundo Devlin *et al*. (2019), o BERT foi criado para pré-treinar representações bidirecionais profundas a partir de um contexto não rotulado, envolvendo conjuntamente tanto o contexto esquerdo como o direito em todas as camadas. Como resultado, o modelo BERT pré-treinado pode ser ajustado com apenas uma camada de saída adicional para criar modelos de última geração que podem ser utilizados para resolver várias tarefas.

Na Figura 17, verifica-se a estrutura geral do BERT, que apresenta duas etapas: 1) pré-treinamento e 2) ajuste fino.

Figura 17 – Estrutura geral do BERT



Fonte: Devlin *et al*. (2019)

No pré-treinamento, o modelo é treinado com dados não rotulados em diferentes tarefas. Já para o ajuste fino, todos os parâmetros do modelo pré-treinado são ajustados usando os dados rotulados da tarefa *downstream*. Dessa forma, cada tarefa possui seu próprio modelo BERT com os parâmetros ajustados para aquela tarefa, porém partindo da mesma inicialização pré-treinada (Devlin *et al*., 2019).

A mesma arquitetura *transformer* é utilizada tanto no pré-treinamento quanto no ajuste fino, sendo alterada apenas a camada de saída conforme a tarefa. *Tokens* especiais como [CLS] e [SEP] são inseridos nas entradas do modelo durante o pré-treinamento para demarcar sentenças e separar pares de sentenças. Portanto, o BERT segue um processo de primeiro pré-treinar uma representação textual genérica antes de ajustá-la para tarefas específicas *downstream* por meio do ajuste fino (Devlin *et al*., 2019).

O modelo inicial foi construído com foco em duas tarefas, sendo a Masked Language Modeling (MLM) e a Next Sentence Prediction (NSP) (Rothman, 2021). A MLM consiste em prever a próxima palavra em uma frase. Na tarefa de pré-treinamento, o objetivo é prever um número de palavras ausentes usando tanto palavras históricas como futuras em uma frase. Ou seja, em uma frase mascara-se um número aleatório de palavras, e o modelo é treinado para prever as palavras que faltam (Ekman, 2021).

Por outro lado, a NSP consiste em prever se duas sentenças estão sequencialmente conectadas no texto original. Sendo assim, o objetivo da NSP é treinar o modelo BERT para entender relacionamentos entre sentenças, o que é crucial para diversas tarefas *downstream* importantes como perguntas e respostas, e inferência de linguagem natural (Devlin *et al*., 2019).

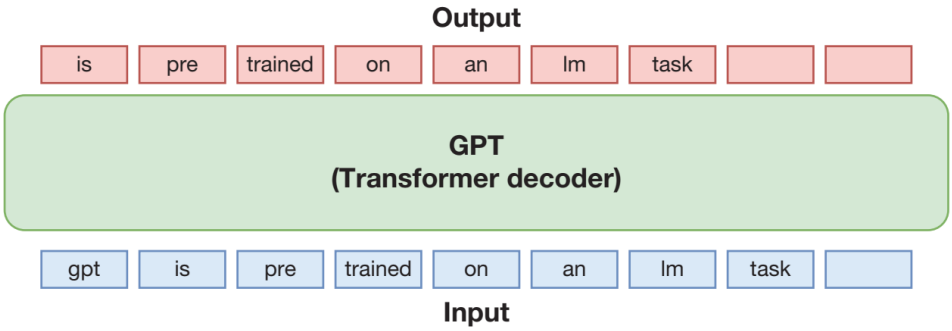
2.4.1.4.3 GPT

Os modelos Generative Pre-Trained Transformers (GPTs) são modelos de linguagem pré-treinados em grandes quantidades de dados textuais que executam muitas tarefas relacionadas à linguagem (Radford *et al*., 2018). Para Ekman (2021), GPT representa um modelo de linguagem natural treinado para prever a próxima palavra, ou seja, a tarefa de pré-treinamento é gerar texto.

O pré-treinamento utiliza a arquitetura do decodificador do transformador (*transformer decoder*), sendo um modelo de autoatenção (seção 2.4.1.4) em que, para determinada palavra, a atenção é calculada utilizando-se apenas as palavras que precedem determinada palavra na sentença frase, de acordo com a ordem de passagem, da esquerda para a direita ou da direita para a esquerda (Radford *et al*., 2018). O mecanismo de autoatenção possibilita ao modelo aprender como diferentes *tokens* se relacionam entre si durante o processo de pré-treinamento.

Na Figura 18, tem-se um exemplo de pré-treinamento do modelo, que apresenta como entrada uma sentença arbitrária. No exemplo, utilizou-se a sentença "*gpt is pre trained on an lm task*". Cada caixa vermelha, na Figura 18, corresponde a uma camada usando uma função de ativação *softmax,* aplicada nos *embeddings* finais para gerar uma distribuição de probabilidade sobre todo o vocabulário. A palavra com maior probabilidade é prevista como sendo a próxima palavra da sentença de entrada. Por exemplo, para a sentença "*gpt is pre trained on an lm task*", o modelo tentaria prever a palavra "*pre"* após observar "*gpt is*" (Ekman, 2021).

Figura 18 – Pré-treinamento do modelo



Fonte: Ekman (2021)

De modo geral, o pré-treinamento é realizado em dados não rotulados e, portanto, pode ser efetuado em grandes quantidades de texto. Após o pré-treinamento, ocorre o ajuste fino para uma tarefa específica, com o auxílio de dados rotulados. As entradas do modelo, bem como a camada de saída, são ligeiramente modificadas para se adequarem melhor à tarefa final para a qual o modelo está sendo usado (Ekman, 2021).

2.5 TRABALHOS CORRELATOS

Esta seção tem o objetivo de evidenciar a contribuição de alguns estudos, mencionados previamente no referencial teórico para fundamentação e elaboração do modelo proposto nesta tese. Na seção 3.3 consta um resumo da revisão integrativa da literatura, com foco na combinação de técnicas de processamento de linguagem natural, representação do conhecimento e aprendizado profundo para auxiliar no processo de classificação de patentes. Os estudos citados a seguir estão disponíveis no Quadro 8 e representam os trabalhos mais aderentes ao contexto desta tese.

Os autores Abdelgawad *et al.* (2019) comparam abordagens recentes para classificação de texto aplicada na tarefa de classificação de patentes. Eles investigaram o quanto uma rede neural pode ser melhorada com uma variedade de métodos de incorporação de palavras diferentes e otimização de hiperparâmetros.

Sofean (2021) apresenta um *pipeline* de aprendizado profundo para classificação automática de patentes com entradas multicanal baseadas em LSTM e incorporação de vetores de palavras.

O estudo dos autores Li *et al*. (2018) apresenta um algoritmo de aprendizado profundo para classificação de patentes baseado em CNN e em incorporação de vetores de palavras, chamado de DeepPatent. Trata-se de um algoritmo que explora a extração automatizada de recursos hierárquicos de CNNs e de modelagem de redes neurais profundas para obter um resultado competitivo na classificação de patentes. Inicialmente tokens de palavras são transformados em vetores de características. Logo após a transformação dos tokens em vetores, os vetores de nível léxico são concatenados para formar a matriz densa de nível de texto. Enquanto isso, os recursos de nível de texto são aprendidos usando uma abordagem convolucional. Por fim, os recursos passam a ser alimentados em uma função sigmoide para prever o rótulo da categoria de patente entre 637 categorias (Li *et al.*, 2018).

O trabalho de Grawe, Martins e Bonfante (2017) tem como objetivo treinar um classificador usando dados textuais de documentos de patente para classificar a patente em uma

das classes da IPC. No contexto da classificação automatizada de patentes, o trabalho propõe testar a eficácia das técnicas atuais de DL em uso para tarefas de processamento de texto, como *Word Embedding* com Word2Vec e DNN com LSTM.

Os autores Zhu *et al.* (2020) apresentaram o método denominado *patent automatic classification method via the symmetric hierarchical convolution neural network* (PAC-HCNN) para classificação automática de patentes. Esse método utiliza a incorporação de palavras para segmentar e vetorizar os dados de entrada. Em seguida, uma rede CNN é utilizada para classificar as patentes com base na incorporação de palavras.

Com o objetivo de investigar a novidade de um pedido de patente, Risch e Krestel (2019) sugerem melhorias na classificação automática de patentes aproveitando técnicas de DL. Os autores treinaram o modelo em um conjunto de dados com mais de 5 milhões de patentes e o avaliaram na tarefa de classificação de patentes. Para isso, propõem uma abordagem de DL baseada em unidades recorrentes com portões (GRUs) para classificação automática de patentes construída com *embeddings* de palavras.

Roudsari *et al.* (2020) utilizam métodos de DL para lidar com o problema de classificação de patentes de subgrupos. Recentemente, métodos pré-treinados de NLP, como o modelo pré-treinado *DistilBERT*, alcançaram resultados promissores em algumas tarefas, principalmente na classificação de texto. O objetivo do trabalho consiste em investigar o efeito da aplicação do modelo pré-treinado *DistilBERT* e ajustá-lo para a tarefa de classificação de patentes multirrótulos.

Min (2021) construiu um sistema de classificação para identificar com precisão os diferentes tipos de inovação em tecnologia na área de energia. Para isso, o modelo de classificação de patente foi construído utilizando uma rede CNN.

A fim de melhorar a eficiência e a precisão da classificação automática de textos de patentes, os autores Lu *et al.* (2019) desenvolveram o modelo C3-BIGRU-AT. Esse modelo promove a integração de redes neurais CNN e BIGRU (do inglês *Bidirectional Gated Recurrent Unit*) com mecanismo de atenção formando um caminho de rede multivariado, incluindo camada de incorporação de palavras, camada de convolução, camada BIGRU, camada de atenção e camada *softmax*.

Huang *et al.* (2020) propõem uma nova estrutura de aprendizagem semissupervisionada de dois estágios, denominada TRIZ-ESSL (*Enhanced Semi-Supervised Learning for TRIZ*). O objetivo principal é aprimorar a classificação automatizada de patentes considerando os conceitos e a estrutura fornecidos pela metodologia TRIZ.

O TRIZ-ESSL utilizou dados não rotulados para treinar um modelo de linguagem recorrente. Logo após o treinamento do modelo, o TRIZ-ESSL inicializa os pesos do modelo baseado em LSTM com o modelo de linguagem recorrente pré-treinado e, em seguida, treina o modelo de classificação de texto em conjuntos rotulados e não rotulados.

A pesquisa de Trappey, Trappey e Hsieh (2021) se concentra na seleção inteligente dos documentos de patentes semanticamente relevantes, utilizando redes neurais com foco em NLP. A incorporação de palavras e de documentos é adotada para representar documentos de patentes em um espaço vetorial, visando permitir a identificação automática de patentes com base na similaridade por meio da métrica do cosseno.

Os autores Liu *et al.* (2021) criaram um algoritmo de recomendação de patentes apoiado na classificação de tópicos e na similaridade semântica. Foram utilizados o título e o resumo de patentes para obter o conjunto de categorias de assuntos através do LLM BERT e métodos de agrupamento DBSCAN. Combina-se esse processo com a estrutura SimNet, que calcula a similaridade entre textos. Obtém-se então um modelo de análise holística treinado para recomendação de patentes.

Os autores Lo e Chu (2021) utilizaram modelos pré-treinados segundo a arquitetura *transformer* para examinar a patenteabilidade de forma automatizada. Os modelos criados têm a capacidade de capturar informações semânticas de reivindicações de patentes para atender os requisitos de patenteabilidade. Os autores empregaram modelos de classificação multirrótulo baseados no BERT-BaseLarge, RoBERTa-BaseLarge e XLNet, com um conjunto de dados da USPTO.

O trabalho de Choi *et al.* (2022) utiliza DL com *transformer* e *graph embedding* para realizar *patent landscaping*, que é a busca por patentes relacionadas a projetos de pesquisa e desenvolvimento. A pesquisa dos autores propõe um modelo de classificação de patentes que combina metadados e informações de texto, utilizando a técnica de *embedding* para redução de dimensões.

Já o trabalho de Jiang *et al.* (2022) propõe um modelo baseado em DL para classificação hierárquica de documentos técnicos na área de engenharia, chamado de TechDoc. Esse modelo usa três tipos de informações: 1) textos em linguagem natural, 2) imagens descritivas em documentos e 3) associações entre documentos. Utiliza-se das redes neurais CNN, RNN e GNN para treinamento do modelo de classificação de documentos com base no IPC de um grande banco de dados de documentos técnicos multimodais.

Por fim, os autores Haghighian Roudsari *et al.* (2022) apresentam um trabalho que investiga o efeito do ajuste fino dos PTMs (BERT, XLNet, RoBERTa e ELECTRA) para a

tarefa essencial de classificação de patentes com vários rótulos. Os modelos foram comparados com abordagens básicas de DL usadas em classificação de patentes. Esses autores utilizaram-se de incorporações de palavras para melhorar o desempenho dos modelos de base.

A partir da pesquisa realizada, conclui-se que nenhum trabalho sobre classificação de patentes apresenta um modelo que envolva, além de técnicas de classificação por meio de redes neurais, elementos de representação de conhecimento, sugestão ordenada (*ranking*) de classes, utilização de grafos de conhecimento para auxiliar na avaliação dos *rankings* e incorporação das decisões dos examinadores no processo de aprendizado.

Portanto, a partir desta tese, pretende-se preencher essa lacuna e apresentar um modelo baseado em métodos e técnicas computacionais e de Engenharia do Conhecimento para auxiliar examinadores no processo de classificação de patentes. Assim, o Quadro 8 contempla as principais características dos trabalhos pesquisados. A última linha explicita as características do modelo proposto. Esse resumo objetiva prover uma visão geral dos conjuntos de dados, métodos de representação de texto e conhecimento, e arquiteturas de rede neural profundas ou LLMs usadas para a tarefa de classificação de patentes.

Quadro 8 – Resumo dos trabalhos relacionados

	Dataset						**Representação**					**Arquitetura**						
Autores	CHINA	CLEF-IP	EPO	USPTO	WIPO	OUTROS	DE	SE	WE	SIMNET	RANK	NLP	MLP	CNN	RNN	LSTM	GRU	BERT
Grawe, Martins e Bonfante (2017)				X					X							X		
Li *et al.* (2018)		X		X					X					X				
Risch e Krestel (2019)		X		X	X				X								X	
Lu *et al.* (2019)						X			X					X			X	
Abdelgawad *et al.* (2020)		X			X				X					X				
Huang *et al.* (2020)	X								X							X		
Roudsari *et al.* (2020)				X					X									X
Zhu *et al.* (2020)	X								X									
Min (2021)						X								X				
Sofean (2021)	X		X						X							X		
Trappey, Trappey e Hsieh (2021)				X			X		X			X						
Liu *et al.* (2021)						X				X								X
Lo e Chu (2021)				X					X									
Haghighian Roudsari *et al.* (2022)		X		X					X					X		X		X
Jiang *et al.* (2022)				X					X					X	X			
Choi *et al.* (2022)						X			X									X
Proposta				X				X			X	X	X	X		X		X

Fonte: elaborado pelo autor (2023)

Referente às legendas do Quadro 8, na dimensão Dataset, CHINA representa patentes chinesas extraídas de algum escritório ou sites, e a coluna OUTROS indica que os autores não especificaram no artigo de onde foram extraídos os conjuntos de dados. Já CLEF-IP é uma coleção de dados descrita na seção 2.2.4, juntamente com o conjunto de dados dos escritórios EPO, USPTO e WIPO. Na dimensão Representação, a sigla DE é a abreviatura de *document embeddings,* utilizada para representar frases, sentenças, parágrafos ou documentos inteiros. A sigla SE é a abreviatura de *sentence embeddings*, usada para representar uma frase ou sentença em um espaço vetorial. Já a sigla WE simboliza os *word embeddings*, todos descritos na seção 2.4.2. O RANK representa uma lista ordenada (*ranking*) de subclasses (neste trabalho, utilizou-se o nível de subclasse) indicando quais subclasses possuem maiores chances de ser atribuídas para determinada patente. Já o SimNet é um modelo para calcular a pontuação de similaridade em textos curtos (Liu *et al*., 2021). Por fim, na dimensão Arquitetura tem-se as soluções apresentadas para resolver o problema da classificação de patentes. As arquiteturas MLP, CNN, RNN e LSTM estão descritas na seção 2.2, e o LLM BERT consta na seção 2.4.1. Um trabalho também utilizou a arquitetura GRU para representar uma rede neural recorrente (Roudsari *et al.*, 2020).

No tocante à representação das patentes, adotou-se a representação via *embeddings* por ser uma abordagem consolidada em NLP para representar semanticamente o texto, permitindo capturar relações entre termos e conceitos. Optou-se por *embeddings* contextualizados pré-treinados, pois eles capturam melhor o significado dos termos considerando determinado contexto.

No que se refere à arquitetura de redes neurais, foram utilizados modelos pré-treinados do tipo BERT por apresentarem bom desempenho em tarefas de NLP, tal como a classificação textual abordada neste tese. Testaram-se diferentes variantes do BERT, mais especificamente por meio do LLM S-BERT, entre elas MiniLM, MPNet e DistilBERT, procurando um balanceamento entre desempenho e custo computacional. Ademais, tal opção teve como objetivo comparar modelos de linguagem mais recentes em relação a arquiteturas de DL mais tradicionais, tais como CNN, LSTM e MLP.

Dessa forma, buscou-se seguir boas práticas consolidadas na literatura quanto às escolhas de *dataset*, representação textual e arquitetura de rede neural para a tarefa de classificação de patentes. Ao mesmo tempo, realizaram-se experimentos com diferentes opções dentro dessas categorias, com vistas a identificar a melhor configuração para a proposta específica desta pesquisa.

2.6 CONSIDERAÇÕES FINAIS

No presente capítulo foram detalhados os conceitos que fundamentam a proposição e o desenvolvimento do modelo deste trabalho.

O conceito de análise de patentes e a tarefa de classificação estão relacionados com os objetivos geral e específicos da tese, e também com a pergunta de pesquisa. A AI, por meio das áreas de processamento de linguagem natural, aprendizado profundo e modelos pré-treinados de redes neurais, forneceu o arcabouço central para o desenvolvimento do modelo de recomendação ordenada de subclasses.

Também foi exposto como explicitar o conhecimento através de métodos e de técnicas de representação do conhecimento, mais especificamente utilizando os conceitos de grafo de conhecimento e representação de conteúdo textual por meio vetores densos (*embeddings*).

Por fim, apresentaram-se os trabalhos correlatos a esta proposta de tese descrevendo-se as principais características. Para a síntese desta seção, foi elaborado um quadro que sumariza os trabalhos em diferentes dimensões, permitindo uma visão geral e deixando clara a contribuição deste trabalho.

3 METODOLOGIA DE PESQUISA

O presente capítulo aborda a metodologia de pesquisa adotada nesta tese. As seções a seguir apresentam o enquadramento metodológico da tese, classificando-a quanto à natureza, aos objetivos e aos procedimentos. Também será apresentada a Design Science Research Methodology, o desenvolvimento da pesquisa e uma síntese da metodologia aplicada.

3.1 ENQUADRAMENTO METODOLÓGICO

Antes de tudo, é preciso discorrer sobre alguns conceitos fundamentais para o desenvolvimento desta pesquisa. Segundo Gil (2008), a ciência tem como objetivo chegar à veracidade dos fatos. Dessa forma, o conhecimento científico se distingue dos demais pelo fato de ter como característica basilar a verificabilidade. Assim, o conhecimento só se torna científico quando estabelecido o método para se chegar a ele. Para Severino (2013, p. 89),

> A ciência utiliza-se de um método que lhe é próprio, o método científico, elemento fundamental do processo do conhecimento realizado pela ciência para diferenciá-la não só do senso comum, mas também das demais modalidades de expressão da subjetividade humana, como a filosofia, a arte, a religião.

Marconi e Lakatos (2003, p. 33) definem método como o

> conjunto das atividades sistemáticas e racionais que, com maior segurança e economia, permite alcançar o objetivo – conhecimentos válidos e verdadeiros –, traçando o caminho a ser seguido, detectando erros e auxiliando as decisões do cientista.

Já a metodologia científica pode ser definida para Gerhardt e Silveira (2009, p. 11) como

> O estudo sistemático e lógico dos métodos empregados nas ciências, seus fundamentos, sua validade e sua relação com as teorias científicas. Em geral, o método científico compreende basicamente um conjunto de dados iniciais e um sistema de operações ordenadas adequado para a formulação de conclusões, de acordo com certos objetivos predeterminados.

Sendo assim, o primeiro passo da pesquisa científica consiste em planejar todas as etapas que serão utilizadas no processo, passando pela escolha do tema, a formulação do problema, a especificação dos objetivos, até chegar à operacionalização dos métodos (Gerhardt; Silveira, 2009).

A pesquisa é a atividade principal da metodologia. Assim, quanto à natureza, esta tese se caracteriza como uma pesquisa aplicada, pois, segundo Gerhardt e Silveira (2009), objetiva gerar conhecimentos para aplicação prática, dirigidos à solução de problemas específicos envolvendo verdades e interesses locais. A pesquisa aplicada tem como particularidade o interesse na aplicação, na utilização e nas consequências práticas dos conhecimentos (Gil, 2008). Ademais, também pode ser vista como aplicada, pois modela parte do conhecimento para uma aplicação real que auxilie pessoas e organizações na tomada de decisão.

Nesse contexto, esta tese ainda pode ser caracterizada como tecnológica, visto que a pesquisa tecnológica tem como objetivo a produção de artefatos e processos. O principal campo de ação da pesquisa tecnológica é a mudança, e os planos gerados para fazer a mudança são considerados inovações, ou seja, cada processo de mudança requer metodologias tecnológicas para gerenciá-lo, tais como novas formas de trabalho (Del Carpio Ramos; Del Carpio Ramos; García-Peñalvo, 2019). Para Cupani (2006), "a tecnologia é uma atividade dirigida à produção de algo novo e não ao descobrimento de algo já existente".

Esta tese adota uma abordagem quantitativa, que é uma forma de testar teorias objetivas através da análise de variáveis mensuráveis, examinando a relação entre elas para que os dados possam ser analisados por meio de procedimentos estatísticos. O relatório final possui uma estrutura fixa composta por introdução, literatura e teoria, métodos, resultados e discussão (Creswell, 2010).

A abordagem quantitativa é caracterizada pela utilização de técnicas padronizadas de coleta de dados, como questionários, testes e escalas de avaliação, e conjuntos de dados para determinada tarefa, e também pela análise estatística dos dados coletados. A coleta de dados é realizada mediante instrumentos de medição que devem representar verdadeiramente as variáveis da pesquisa. As respostas obtidas são codificadas e transferidas para uma matriz de dados e preparadas para análise. Alguns tipos de instrumentos que podem ser empregados em coleta de dados são questionários, escalas de mensuração de atitudes, análise de conteúdo quantitativo, observação, testes padronizados e inventários (Sampieri; Collado; Lucio, 2013).

Em relação aos objetivos, a tese se enquadra como exploratória, pois tem pretende proporcionar uma visão geral, de tipo aproximativo, acerca de determinado fato. Realiza-se esse tipo de pesquisa especialmente quando o tema escolhido é pouco explorado, tornando-se difícil formular hipóteses precisas e operacionalizáveis sobre ele (Gil, 2008).

Segundo Gil (2008), parte dos estudos exploratórios pode ser definida como pesquisas bibliográficas. Em relação aos procedimentos, esta tese se caracteriza como uma pesquisa bibliográfica, pois se desenvolve com base no acervo de documentos já existentes, composto

principalmente de artigos científicos e livros. A pesquisa bibliográfica tem como vantagem primordial permitir ao pesquisador a cobertura de uma série de fenômenos muito mais ampla do que aquela que poderia pesquisar diretamente (Gil, 2008).

Outra característica deste trabalho é a relação com a *Design Science* (DS). Simon (1996) cunhou o termo "ciência do artificial", que diz respeito à "concepção de artefatos que realizem objetivos". É considerado o precursor da *Design Science* (DS), como uma ciência do desenvolvimento de artefatos (Bax, 2013). Segundo Simon (1996), a DS é um paradigma pragmático de pesquisa que busca criar artefatos inovadores para resolver problemas do mundo real.

A *Design Science* destina-se a criar o conhecimento, e não apenas aplicá-lo, ou seja, é uma ciência que está concentrada no *design,* sendo voltada para desenvolver e projetar soluções que aperfeiçoem os sistemas existentes, resolver problemas ou mesmo criar artefatos que contribuam para um melhor comportamento humano, seja na sociedade, seja nas organizações (Dresch; Lacerda; Antunes Jr., 2015).

Dessa forma, a DS é considerada a base para a *Desing Science Research* (DSR), que representa o método adotado para a sistematização desta tese. Segundo Bax (2013), o método DSR é uma estratégia de pesquisa capaz de orientar tanto a construção do conhecimento quanto aprimorar as práticas em sistemas de informação e de várias disciplinas relacionadas ao campo gerencial e tecnológico da ciência da informação, indo de acordo com a linha de pesquisa Teoria e Prática em Engenharia do Conhecimento.

No Quadro 9, tem-se uma síntese da classificação desta pesquisa quanto a sua natureza, abordagem, objetivo e procedimentos metodológicos.

Quadro 9 – Síntese da classificação da pesquisa da tese

Tipo de pesquisa	**Classificação**
Natureza	Tecnológica
Abordagem	Quantitativa
Objetivo	Exploratória
Procedimentos	Bibliográfica

Fonte: elaborado pelo autor (2023)

Diante do exposto, apresentam-se na próxima seção os princípios e os fundamentos da DSR que conduzem para determinada metodologia a ser seguida no desenvolvimento desta pesquisa.

3.2 DESIGN SCIENCE RESEARCH METHODOLOGY

Nos últimos anos, a Design Science Research (DSR) foi bastante difundida, em especial nas áreas de sistemas de informação e engenharias. Segundo Hevner (2020), a DSR é um paradigma, também entendido como método, fundamentado na solução de problemas para aprimorar o conhecimento humano com a criação de artefatos inovadores.

A DSR implica em construir, investigar, validar e avaliar artefatos com o propósito de resolver novos problemas práticos, tais como constructos, *frameworks*, modelos, métodos e instâncias de sistema de informação. Engloba também o estudo de métodos, comportamentos e melhores práticas relacionadas com a análise do problema e com o processo de desenvolvimento de artefato são abrangidos (Bax, 2013).

Os artefatos podem conter construções, modelos, métodos e instanciações (Hevner *et al.*, 2004), assim como incluir inovações sociais (Aken, 2004) ou novas propriedades de recursos técnicos, sociais ou informacionais (Järvinen, 2007). Para Peffers *et al.* (2007), essa definição compreende qualquer objeto projetado com uma solução embutida para um problema de pesquisa.

Segundo Hevner *et al.* (2004), os resultados da DSR incluem tanto os artefatos recém-projetados (por exemplo, dispositivos, protocolos e sistemas) quanto uma compreensão mais completa do motivo pelo qual esses artefatos fornecem um aprimoramento, ou mesmo uma interrupção, para os contextos de aplicativos relevantes.

A criação de artefatos de sucesso está fortemente relacionada à condução de pesquisas em Design Science (DS) em Sistemas de Informação (SI). Porém, a carência de uma metodologia que seja aceita e sirva de *framework* para pesquisas em DS e de um modelo para sua apresentação pode ter contribuído para a criação da Design Science Research Methodology (DSRM). A DSRM incorpora princípios, práticas e procedimentos necessários para realizar essas pesquisas e atende três objetivos: a) ser consistente com a literatura prévia; b) prover orientação para pesquisadores; e c) fornecer um modelo mental para a apresentação dos resultados (Peffers *et al*., 2007).

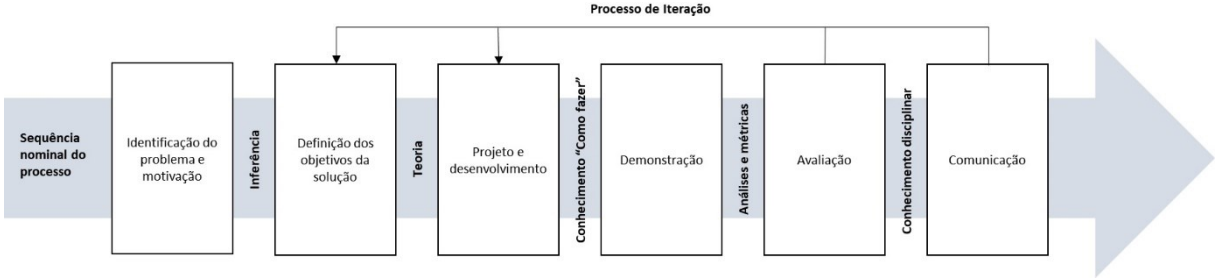
A DSRM cria e avalia artefatos destinados a resolver problemas organizacionais identificados (Gregório *et al.*, 2021). Os autores Hevner *et al.* (2004) estipularam regras para a DSRM, devendo a metodologia ser baseada em conhecimentos e teorias existentes, de forma a produzir uma solução relevante (ou seja, o artefato) para um problema específico em um determinado domínio que deverá ter sua utilidade, qualidade e eficácia avaliadas com rigor. Por

fim, o ciclo deve terminar com a comunicação das descobertas relevantes e do conhecimento disciplinar para as principais partes interessadas tanto na academia quanto na prática.

A DSRM envolve um modelo de processo rigoroso que consiste em seis atividades (Peffers *et al.*, 2007), as quais serão utilizadas nesta tese, conforme ilustra a Figura 19. As atividades ocorrem de forma sequencial ou de acordo com a necessidade do projeto. A seguir, descreve-se brevemente cada uma dessas atividades:

- Atividade 1 - Identificação do problema e motivação: esta atividade é responsável pela definição do problema de pesquisa específico, assim como pela justificativa da importância da solução proposta;
- Atividade 2 - Definição dos objetivos da solução: esta atividade deduz os objetivos de uma solução a partir da definição do problema e do conhecimento do que é possível e viável executar;
- Atividade 3 - Projeto e desenvolvimento: esta atividade trata da criação do artefato. Inclui a determinação das funcionalidades desejadas do artefato, sua arquitetura e, em seguida, a criação do artefato real;
- Atividade 4 - Demonstração: esta atividade se preocupa com a demonstração do uso do artefato para resolver o problema. Isso implica no uso do artefato em experimentação, simulação, estudo de caso, prova ou outra atividade apropriada;
- Atividade 5 - Avaliação: esta atividade refere-se à observação e à medição do artefato para a solução do problema. Isso implica na comparação dos objetivos de uma solução com os resultados reais observados do uso do artefato na demonstração;
- Atividade 6 - Comunicação: esta atividade se destaca pela divulgação do problema e sua importância, assim como a apresentação do artefato produzido.

Figura 19 – Atividades para a condução da DSRM



Fonte: adaptado de Peffers *et al.* (2007, p. 54)

Nesta pesquisa será utilizada a Design Science Research Methodology (DSRM) para propor um modelo voltado à tarefa de classificação de patentes a partir de fontes de dados não estruturados na forma de texto, levando em conta aspectos de ordenação de subclasses e explicitação de conhecimento. Para isso, será empregado o método proposto por Peffers *et al.* (2007), o qual evidencia os passos utilizados para a solução do problema.

3.3 REVISÃO INTEGRATIVA DA LITERATURA

Na presente pesquisa propõe-se um modelo que estabelece uma combinação de técnicas de processamento de linguagem natural, representação do conhecimento e aprendizado profundo para auxiliar a classificação de patentes. Mais especificamente, pretende-se contribuir com o processo de tomada de decisão, ou seja, facilitar a tarefa de classificação de patentes por examinadores gerando recomendações ordenadas (*ranking*) de subclasses, assim como provendo meios que auxiliem no entendimento das sugestões.

Para identificar o problema e propor uma solução, é fundamental realizar buscas na literatura com o intuito de apontar uma lacuna de conhecimento em um domínio específico.

Este estudo se baseou em uma revisão integrativa da literatura, método específico de pesquisa que visa delinear uma análise sobre o conhecimento já levantado em pesquisas anteriores sobre um determinado tema (Botelho; Cunha; Macedo, 2011). Assim, foi realizada uma revisão integrativa da literatura para obter melhor compreensão sobre o tema proposto. Dessa forma, procedeu-se a uma busca na literatura através da expressão *(“patent classification” OR “patent document classification” OR “patent text classification” OR “patent document categorization”) AND (“machine learning” OR “artificial intelligence” OR “deep learning” OR “neural network” OR “knowledge graph”)* nas bases de dados Science Direct®, Scopus®, Web of Science® (WoS) e IEEExplore®. Essas bases de dados foram escolhidas por possuir credibilidade e relevância no âmbito acadêmico.

A revisão teve início em 2021 e passou por atualização em 2022, sempre buscando por novos artigos relacionados ao tema de pesquisa. O período estipulado para a busca foi de 2017 a 2022, preferencialmente para artigos publicados nos últimos seis anos, sendo a última busca realizada em junho de 2022. A escolha desse período permitiu uma análise atualizada e focada nas publicações mais recentes sobre o tema. Com isso, garante-se a inclusão de estudos representativos do estado da arte e que empregam técnicas modernas de Inteligência Artificial e aprendizado de máquina para a classificação de patentes. Dessa forma, o período estipulado

adequa-se aos objetivos de analisar o conhecimento recente na área e identificar oportunidades de pesquisa dentro de um escopo factível.

A estratégia de busca, além das palavras contidas na expressão de busca e nos operadores booleanos, levou em consideração o local onde as palavras da expressão de busca poderiam ser encontradas. Ponderou-se encontrar as palavras da expressão de busca no título, no resumo e nas palavras-chave do documento. Sendo assim, a Science Direct® retornou 113 documentos, a Scopus® 48 documentos, e a Web of Science® 33 documentos, contabilizando um total de 194 documentos.

Com a atualização em julho de 2022, foram incluídos à lista inicial 5 documentos referentes à pesquisa realizada com a expressão utilizada anteriormente, acrescentando um limite de busca para as publicações entre 2021 e 2022. A base Science Direct® retornou 8 documentos, a base IEEExplore® retornou 14 documentos, a base Scopus® retornou 23 documentos, e na Web of Science® foram retornados 13 documentos, contabilizando um total de 58 documentos. Desse total, retirando-se os duplicados de 2021 e realizando-se a leitura do título e do resumo, selecionaram-se 5 documentos condizentes com o tema desta pesquisa.

O critério para exclusão foi utilizado a fim de remover documentos duplicados encontrados em mais de uma base de dados, ou seja, permanecendo somente um documento. Após a adoção do critério para exclusão, incluindo a nova pesquisa de 2022, restou um total de 64 documentos para análise. Foram considerados somente os documentos que estavam de acordo com a busca realizada e com a questão de pesquisa.

Os dados foram extraídos após a leitura dos títulos, dos resumos e das palavras-chave de todas as publicações completas, sendo identificados 33 documentos. Em seguida, esses documentos foram relacionados em uma matriz de síntese para análise e adequação aos critérios para inclusão da pesquisa. Os critérios para inclusão de documentos utilizados foram:

- os termos de busca devem constar no título, no resumo ou na palavra-chave;
- os documentos devem estar preferencialmente no idioma inglês;
- os documentos publicados devem respeitar o período de 2017 a 2022; e
- os documentos devem estar disponíveis para download.

Analisando-se os 33 documentos resultantes da revisão integrativa da literatura (Quadro 10), percebe-se que nenhum dos estudos relacionados investigou a classificação de patentes associada à integração de elementos que explicitem o conhecimento no cenário de análise e gestão de patentes com o intuito de auxiliar no processo de tomada de decisão.

Quadro 10 – Áreas, métodos, técnicas e algoritmos utilizados na classificação de patentes

(continua)

id	Título	Autor	Área	Método, técnica e algoritmo
1	A New Function-Based Patent Knowledge Retrieval Tool for Conceptual Design of Innovative Products	(Liu *et al*., 2020)	Aprendizado de máquina	Naïve Bayes
2	An Extension-Based Classification System of Cloud Computing Patents	(Huang; Tan, 2020)	Aprendizado de máquina	Ontologia, Gray Relational Analysis (GRA) e Singular Value Decomposition (SVD)
3	Automated Classification of Patents: A Topic Modeling Approach	(Yun; Geum, 2020)	Aprendizado de máquina	LDA, SVM
4	Classification of Patents according to Industry 4.0 Pillars using Machine Learning Algorithms	(Jafery *et al*., 2019)	Aprendizado de máquina	SVM
5	Multi-label classification and interactive NLP-based visualization of electric vehicle patent data	(De Clercq *et al*., 2019)	Aprendizado de máquina	Árvore de decisão, Floresta aleatória e KNN
6	On the Potential of Taxonomic Graphs to Improve Applicability and Performance for the Classification of Biomedical Patents	(Frerich *et al*., 2021)	Aprendizado de máquina	SVM, KNN, ANN, LogReg
7	Parameter tuning Naïve Bayes for automatic patent classification	(Cassidy, 2020)	Aprendizado de máquina	Naïve Bayes
8	Patent Classification via Textual Analysis Which Sections to be Included?	(Yücesoy Kahraman; Dereli; Durmuşoğlu, 2018)	Aprendizado de máquina	SVM, k-NN
9	Patent Text Classification Based on Naive Bayesian Method	(Xiao; Wang; Liu, 2018)	Aprendizado de máquina	Naïve Bayes
10	Research on Patent Text Classification Based on Word2Vec and LSTM	(Xiao; Wang; Zuo, 2018)	Aprendizado de máquina	KNN, Word2Vec, LSTM, CNN
11	Automatic Patents Classification Using Supervised Machine Learning	(Shahid *et al*., 2020)	Aprendizado de máquina, Rede neural	Naïve Bayes, KNN, ANN, SVM, Ada BoostM1

(continuação)

id	Título	Autor	Área	Método, técnica e algoritmo
12	An Intelligent Patent Recommender Adopting Machine Learning Approach for Natural Language Processing: a Case Study for Smart Machinery Technology Mining	(Trappey; Trappey; Hsieh, 2021)	Aprendizado de máquina, Rede neural	doc2vec
13	A Structured Representation Framework for TRIZ-Based Chinese Patent Classification via Reinforcement Learning	(Yu *et al*., 2020)	Aprendizado por reforço, Rede neural	SVM, LSTM e BERT
14	A Feasible Dashboard to predict Patent Mining Using Classification Algorithms	(Naik; Brunda; Seema, 2020)	Data mining	Árvore de decisão; Naïve Bayes
15	A Patent Text Classification Model Based on Multivariate Neural Network Fusion	(Lu *et al*., 2019)	Rede neural	CNN, BIGRU
16	A Semi-Supervised Learning Framework for TRIZ-Based Chinese Patent Classification	(Huang *et al*., 2020)	Rede neural	LSTM
17	Patent Automatic Classification Based on Symmetric Hierarchical Convolution Neural Network	(Zhu *et al*., 2020)	Rede neural	CNN
18	A Hierarchical Feature Extraction Model for Multi-Label Mechanical Patent Classification	(Hu *et al*., 2018)	Rede neural	HFEM, CNN, LSTM e BiLSTN
19	Domain-Specific Word Embeddings for Patent Classification	(Risch; Krestel, 2019)	Aprendizado profundo, Rede neural	bi-directional gated recurrent units (GRUs), RNN patent
20	Early Detection of Valuable Patents Using a Deep Learning Model: Case of Semiconductor Industry	(Chung; Sohn, 2020)	Aprendizado profundo, Rede neural	CNN e Bi-LSTM
21	MEXN: Multi-Stage Extraction Network for Patent Document Classification	(Bai; Shim; Park, 2020)	Aprendizado profundo, Rede neural	CNN e RNN patent
22	Optimizing Neural Networks for Patent Classification	(Abdelgawad *et al*., 2019)	Aprendizado profundo, Rede neural	CNN
23	DeepPatent: Patent Classification with Convolutional Neural Networks and Word Embedding	(Li *et al*., 2018)	Aprendizado profundo, Rede neural	CNN

(conclusão)

id	Título	Autor	Área	Método, técnica e algoritmo
24	Deep Learning Based Pipeline with Multichannel Inputs for Patent Classification	(Sofean, 2021)	Aprendizado profundo, Rede neural	LSTM
25	Automated Patent Classification Using Word Embedding	(Grawe; Martins; Bonfante, 2017)	Aprendizado profundo	word2vec, RNN, LSTM
26	Multi-label Patent Classification using Attention-Aware Deep Learning Model	(Roudsari *et al*., 2020)	Aprendizado profundo	PNL, DistiBERT, BiLSTM
27	Power Patent Classification Method Based on Deep Neural Network	(Min, 2021)	Aprendizado profundo	CNN
28	A Patent Recommendation Algorithm Based on Topic Classification and Semantic Similarity	(Liu *et al*., 2021)	Aprendizado profundo	TF-IDF, BERT, DBSCAN e SimNet
29	Pre-trained Transformer-based Classification for Automated Patentability Examination	(Lo; Chu, 2021)	Aprendizado profundo	BERT-Base, BERT-Large, RoBERTa-Base, RoBERTa-Large e XLNet
30	Leveraging Label Hierarchy Using Transfer and Multi-Task Learning: A Case Study on Patent Classification	(Aroyehun *et al*., 2021)	Aprendizado profundo	GRU (Gated Recurrent Unit); RNN-patent
31	PatentNet: Multi-Label Classification of Patent Documents Using Deep Learning Based Language Understanding	(Haghighian Roudsari *et al*., 2022)	Aprendizado profundo	BERT, XLNet, RoBERTa e ELECTRA e LSTM, BiLSTM, CNN e CNN-BiLSTM
32	Deep Learning for Technical Document Classification	(Jiang *et al*., 2022)	Aprendizado profundo	CNN, RNN e GNN
33	Deep Learning for Patent Landscaping Using Transformer and Graph Embedding	(Choi *et al*., 2022)	Aprendizado profundo	Transformer (TRF), Diff2Vec, APL e classificador baseado emmBERT (PATENTBERT)

Fonte: elaborado pelo autor (2023)

O Quadro 10 está dividido em oito áreas encontradas na revisão da literatura. As linhas de 1 a 10 correspondem à aplicação de ML para a classificação de patentes. As linhas 11, 12 e 13 se referem à utilização de ML em conjunto com redes neurais. O trabalho relacionado na linha 14 utiliza-se de DM por meio de técnicas de classificação. As linhas 15 a 18 referem-se à utilização de redes neurais, enquanto as linhas 19 a 24 combinam técnicas de aprendizado profundo com redes neurais. Já as linhas 25 a 33 listam trabalhos baseados em DL como técnica para classificar patentes.

A extração dessas técnicas permitiu a definição inicial do problema de pesquisa e de algumas possibilidades para solucioná-lo. Quanto ao ineditismo da tese, verificou-se que as pesquisas relacionadas à classificação de patentes, em sua maioria, descrevem os métodos e/ou modelos utilizados no nível de classificação de classe e subclasse de patentes. O foco da tese reside nas subclasses.

Percebe-se o esforço nas pesquisas para classificar patentes de forma automática e com maior precisão, combinando várias técnicas e algoritmos que auxiliem na busca e na recuperação de documentos de patentes. Os métodos e as técnicas no contexto de NLP se mostram eficientes na tarefa de classificar patentes, mesmo que não se tenha um procedimento único para tal.

A revisão da literatura aponta que os métodos e as técnicas computacionais e de Engenharia do Conhecimento são fundamentais para a tomada de decisão, auxiliando examinadores na classificação adequada de patentes. Ou seja, ao chegar uma nova patente, o ferramental ofertado ao examinador deve prover suporte adequado para a realização da tarefa de classificar essa patente.

Sendo assim, a proposição e o desenvolvimento de um modelo que auxilie na tarefa de classificação de patentes fazem-se necessários para permitir uma classificação mais assertiva, minimizando custos e recursos durante o processo de tomada de decisão. Nesse contexto, identificou-se uma lacuna que trata do tema de forma inédita, consistindo na proposição de um modelo que objetiva a recomendação de subclasses de patentes a partir de fontes de dados não estruturados na forma de texto levando-se em conta aspectos de ordenação de classes (*ranking*) e explicitação de conhecimento.

3.4 DESENVOLVIMENTO DA PESQUISA

Esta pesquisa é de natureza aplicada e tecnológica com objetivos exploratórios, pois possui como enfoque principal a proposição de um modelo voltado à classificação de patentes

a partir de fontes de dados não estruturados na forma de texto, levando em conta aspectos de ordenação de subclasses e explicitação de conhecimento. Para tal, será utilizada a Design Science Research Methodology (DSRM) proposta por Peffers *et al.* (2007), evidenciando os passos utilizados para a solução do problema (Quadro 11).

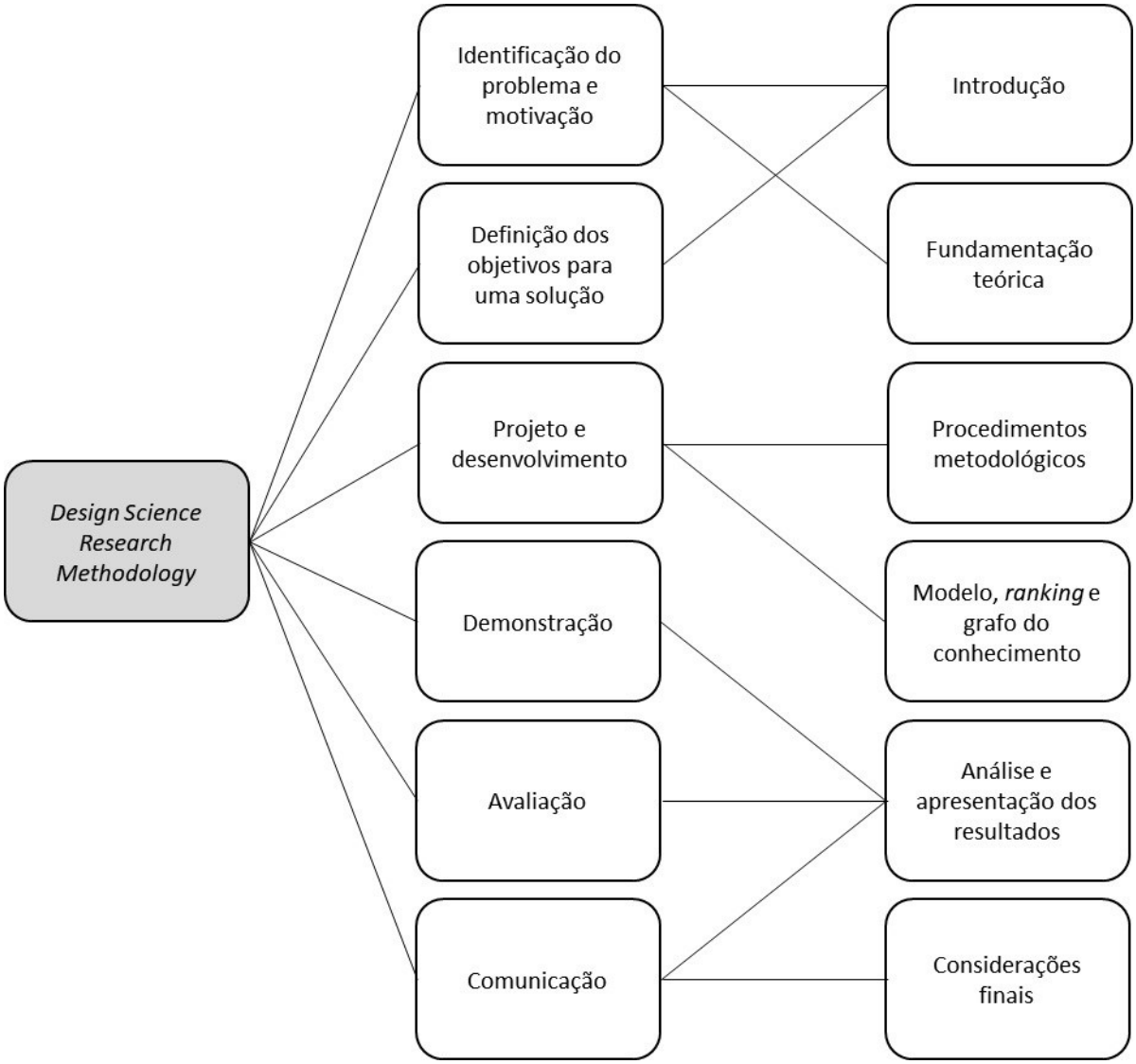
Quadro 11 – Atividades de pesquisa

Atividades	**Descrição**
Identificação do problema e motivação	- Como auxiliar na análise de patentes, mais especificamente na tarefa de classificação, por meio de elementos que caracterizem a relevância de determinada categoria e explicitem o conhecimento latente presente em bases de patentes?
Definição dos objetivos para uma solução	- Propor um modelo voltado à classificação de patentes a partir de texto levando em conta aspectos de ordenação de subclasses e explicitação de conhecimento. - Identificar métodos e técnicas que possibilitem sugerir subclasses de maneira ordenada pela sua relevância. - Analisar métodos e técnicas que permitam explicitar conhecimento latente em bases de patentes. - Avaliar a viabilidade do modelo proposto por meio do desenvolvimento de um sistema (nível de protótipo) considerando cenários de estudo.
Projeto e desenvolvimento	- Modelo voltado à classificação de patentes a partir de fontes de dados não estruturados na forma de texto, levando em conta aspectos de ordenação de subclasses e explicitação de conhecimento.
Demonstração	- Demonstração do modelo utilizando cenários de estudo.
Avaliação	- Avaliar a aplicabilidade do modelo no cenário criado. - Realizar testes com arquiteturas de DNN. - Discutir o impacto do *ranking* de subclasses e da explicitação do conhecimento por meio dos grafos de conhecimento.
Comunicação	- Comunicação à comunidade acadêmica dos resultados obtidos.

Fonte: elaborado pelo autor (2023)

Na Figura 20, listam-se as atividades da DSRM relacionadas com os capítulos da tese.

Figura 20 – Atividades de desenvolvimento da tese



Fonte: adaptado de Peffers *et al.* (2007)

Nas próximas seções, são detalhadas as etapas do desenvolvimento da tese em conformidade com as atividades da DSRM.

3.4.1 Definição do problema e motivação

Os primeiros passos para o início da tese estiveram relacionados à identificação de um problema e explicitação da motivação do trabalho. Para tal, foi realizada uma revisão integrativa da literatura (seção 3.3) com protocolo disponível no Apêndice B, por meio da qual se procurou encontrar uma lacuna de pesquisa no contexto da Engenharia do Conhecimento. Após a pesquisa, verificou-se a falta de trabalhos com foco na recomendação de subclasses de patentes

de maneira ordenada, assim como na explicitação do conhecimento dessas sugestões, visando auxiliar na tomada de decisão por parte de examinadores.

No total, foram levantados 244 documentos, pesquisados nas bases Science Direct®, Scopus®, Web of Science® e IEEExplore®. Desses documentos, 33 foram selecionados para compor a matriz de síntese.

3.4.2 Definição dos objetivos

Com base na revisão integrativa da literatura para a definição do problema e motivação, foram delineados os objetivos desta pesquisa – geral e específicos – que promovem suporte ao modelo proposto na tese e a sua avaliação por meio de um cenário de estudo, assim como a formulação da pergunta de pesquisa.

3.4.3 Projeto e desenvolvimento

O principal entregável desta atividade deve ser o artefato, nesta tese representado pelo modelo proposto, de forma que permita, como já mencionado, recomendar um conjunto de subclasses de patentes, bem como explicitar o conhecimento envolto nas sugestões, visando subsidiar a tomada de decisão por parte de determinado examinador na tarefa de classificação de patentes. Nesse sentido, esta tese lança mão de um conjunto de métodos e técnicas computacionais e de Engenharia do Conhecimento, buscando prover meios de atender adequadamente demandas da GC, mais especificamente no cenário de análise e gestão de patentes. As atividades para o desenvolvimento do artefato (modelo) são apresentadas nas seções a seguir.

3.4.3.1 Coleta dos dados

Os dados de patentes utilizados para a avaliação do modelo foram selecionados da base de patentes americanas USPTO®. A escolha ocorreu pelo acesso facilitado aos dados de patentes e, também, pela expressiva representatividade dessa base no mercado tecnológico.

Particularmente, esta tese utiliza o conjunto de dados USPTO-2M[23], que é uma referência para a tarefa de classificação de patentes. Esse conjunto de dados fornece 2.000.147

[23] Disponível em: http://mleg.cse.sc.edu/DeepPatent/index.html e https://github.com/JasonHoou/USPTO-2M. Acesso em: 12 out. 2021.

registros (patentes) após a limpeza dos dados, de um total de 2.679.443 documentos brutos de patentes de utilidade dos EUA, contendo 637 categorias ao nível de subclasse (Li *et al*., 2018).

Ao todo, são disponibilizadas patentes de 2006 a 2015 separadas por ano. Por exemplo, o arquivo 2014_USPTO.JSON refere-se às patentes extraídas do ano de 2014. Os arquivos contendo os dados de patentes possuem a mesma estrutura, como se segue: rótulo de subclasse (*subclass_labels*), indicando uma ou mais subclasses; resumo (*abstract*); título (*title*) e número (*no*), como apresentado na Figura 21. Além desses dados, incluiu-se nos documentos de patentes o ano da extração, correspondendo ao ano em que o conjunto de dados foi extraído, ou seja, todos os documentos do arquivo 2014_USPTO.JSON receberam o ano 2014.

Figura 21 – Exemplo do arquivo 2014_USPTO.JSON


```
[
    {
        "Subclass_labels": [
            "A61B",
            "G09B"
        ],
        "Abstract": "a method is presented to address quantitative assessment of facial emotion sensitivity of a subject where the meth
        "Title": "method and system for quantitative assessment of facial emotion sensitivity",
        "No": "US08777630"
    },
    {
        "Subclass_labels": [
            "E21B"
        ],
        "Abstract": "in a method for drilling a borehole real time geosteering data including natural gamma ray data is obtained for a
        "Title": "method for drilling a borehole",
        "No": "US08857538"
    },
    {
        "Subclass_labels": [
            "C23C",
            "H01J"
        ],
        "Abstract": "a magnetron actuator for moving a magnetron in a nearly arbitrary radial and azimuthal path in the back of a targe
        "Title": "homing device for magnetron rotating on two arms",
        "No": "US08900427"
    },
```


Fonte: elaborado pelo autor (2023)

Cabe mencionar que a base de dados é entendida como a base de conhecimento, uma vez que também armazena a representação vetorial densa (*embedding*) de cada patente. Na indexação, removeu-se um total de 1.739 patentes por não apresentarem algum dos campos de armazenamento.

3.4.3.2 Pré-processamento dos dados

Como mencionado, o conjunto de dados USPTO-2M é uma referência para a classificação de patentes, contendo patentes de 2006 a 2015, sendo os dados organizados no formato JSON. De maneira geral, cada patente é disponibilizada sem caracteres de pontuação,

ou seja, um processamento básico de limpeza dos dados já foi realizado. As principais ferramentas aplicadas na limpeza dos dados estão descritas na seção 2.2.

A partir do dado original, realizou-se um pré-processamento, o qual consiste na transformação do texto que servirá de entrada para o modelo. Essa transformação considera a concatenação do título e resumo, a conversão do texto para letras minúsculas e a remoção das *stopwords*. Na base de conhecimento foram produzidos dois esquemas para a avaliação, um contendo os dados das patentes com *stopwords* e o outro esquema sem *stopwords,* sendo esses esquemas utilizados para a avaliação do modelo.

3.4.3.3 Transformação dos dados

A incorporação de documentos (*document embeddings*) apresentada na seção 2.3.2 é constituída de técnicas de NLP usadas para mapear um documento no espaço vetorial. O vetor resultante de palavras, frases e documentos quantifica os dados de texto não estruturados, permitindo que métodos de análise quantitativa possam ser usados (Mikolov *et al.*, 2021a).

Após o pré-processamento dos dados, o passo seguinte consiste na geração da representação vetorial densa (*embedding*) das patentes para cada um dos esquemas discutidos na seção anterior. Uma vez que o *embedding* é gerado, este é armazenado na base de conhecimento, o que possibilita a realização de consultas aproximadas calculando-se a similaridade vetorial. Para tal, a similaridade de cosseno é utilizada visando encontrar, para determinada patente de entrada, as patentes mais semelhantes.

Vale destacar que para a geração de *embeddings* um PTM deve ser utilizado objetivando mapear o texto de patentes para um espaço vetorial denso. Ressalta-se que cada modelo pré-treinado possui uma dimensionalidade específica que deve ser considerada no momento em que o vetor é armazenado na base de conhecimento.

Por fim, parte dos vetores densos de patentes servirá para a etapa de treinamento (por meio da composição da base de conhecimento) e parte servirá para o teste do modelo, nesse caso em particular entendido como o cenário geral. Já para o cenário específico, parte dos vetores densos servirá de instância para o treinamento de algumas arquiteturas de redes neurais e parte servirá para a realização dos testes.

3.4.3.4 Treinamento

Para esta etapa, o conjunto de dados é separado em duas partes, uma para o treinamento e outra para o teste. Na Figura 22, são apresentados os campos com os dados extraídos de uma patente do conjunto de dados USPTO-2M, que foi indexada para produzir os testes no modelo.

Figura 22 – Exemplo de uma patente


```
{
    "Subclass_labels": ["A47D"],
    "Abstract": "a baby crib includes a frame structure a support base located at a bottom of the frame
    the support base including at least one opening and a cushion pad disposed on the support base the
    cushion pad including at least one strap disposed at a position corresponding to the opening the strap
    has a first end secured with the cushion pad and a second end that is formed as a free end adapted to
    pass through the opening and detachably fasten with the frame structure via a fastener as the cushion
    pad is securely held with the support base of the baby crib injuries caused by accidental lift of
    the cushion pad can be prevented",
    "Title": "baby crib",
    "No": "US08925127"
}
```


Fonte: patente obtida a partir do conjunto de dados USPTO-2M

O conjunto de dados de treinamento voltado à criação da base de conhecimento é composto por patentes de 2006 a 2014. Como já mencionado anteriormente, cada patente destinada ao treinamento é armazenada na base de conhecimento contendo as informações originais acrescidas da representação vetorial densa (*embedding*) da patente. Nessa etapa, foram utilizados dois PTMs, sendo um contextual e o outro não contextual. Já o conjunto de testes é representado pelas patentes do ano de 2015.

Para a comparação do modelo proposto com as redes de DL (CNN, LSTM e MLP), o conjunto de dados foi reduzido considerando as patentes dos anos de 2012, 2013 e 2014. Cada uma das patentes foi armazenada em uma tabela contendo o título, o resumo, a lista de classes e um tipo – este último para representar se a patente será utilizada na fase de treinamento ou de teste.

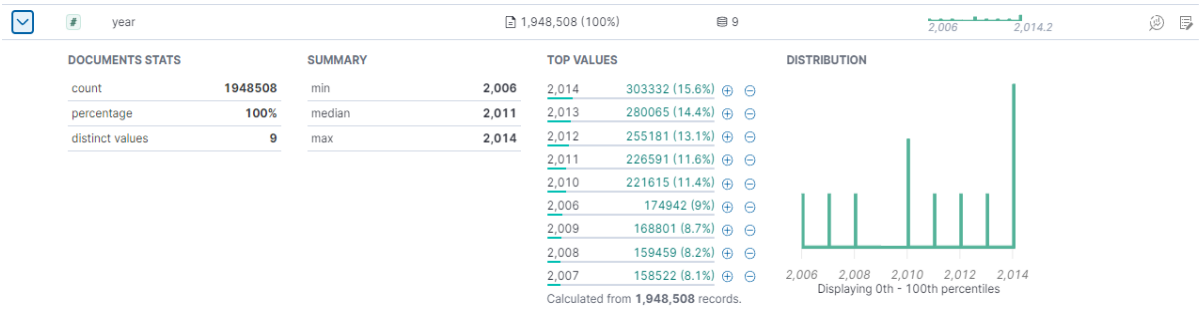
Para o cenário de estudo, é realizada a análise de todas as patentes de treinamento e teste para aferição do modelo, com um total de 1.998.408 (um milhão, novecentas e noventa e oito mil, quatrocentas e oito) patentes. O total de patentes foi utilizado para o cenário geral de avaliação do modelo. Além disso, definiu-se um cenário específico para a comparação do modelo proposto com arquiteturas de redes neurais (CNN, LSTM e MLP) utilizando-se 50 mil patentes.

No intuito de se avaliar o modelo, foram consideradas as patentes dos anos de 2006 a 2014 (Figura 23) para compor o conjunto de treinamento, totalizando 1.948.508 patentes indexadas e armazenadas em uma base de dados contendo o número, o título, o resumo, a subclasse e o ano da patente. As patentes de 2015 são usadas para compor o conjunto de teste

do modelo, com um total de 49.900 patentes. As Figuras 23, 24 e 25 representam os dados armazenados no Elasticsearch®, e a ferramenta Kibana® foi utilizada para exploração e visualização dos dados armazenados.

Percebe-se o aumento do número de patentes com o passar dos anos – 2006 com um total de 174.942 patentes indexadas e 2014 com 303.332 patentes indexadas, com média de 216.501 patentes indexadas no período de 2006 a 2014. Trata-se de uma base de dados extensa com aproximadamente 2 milhões de patentes entre o conjunto de treinamento e teste.

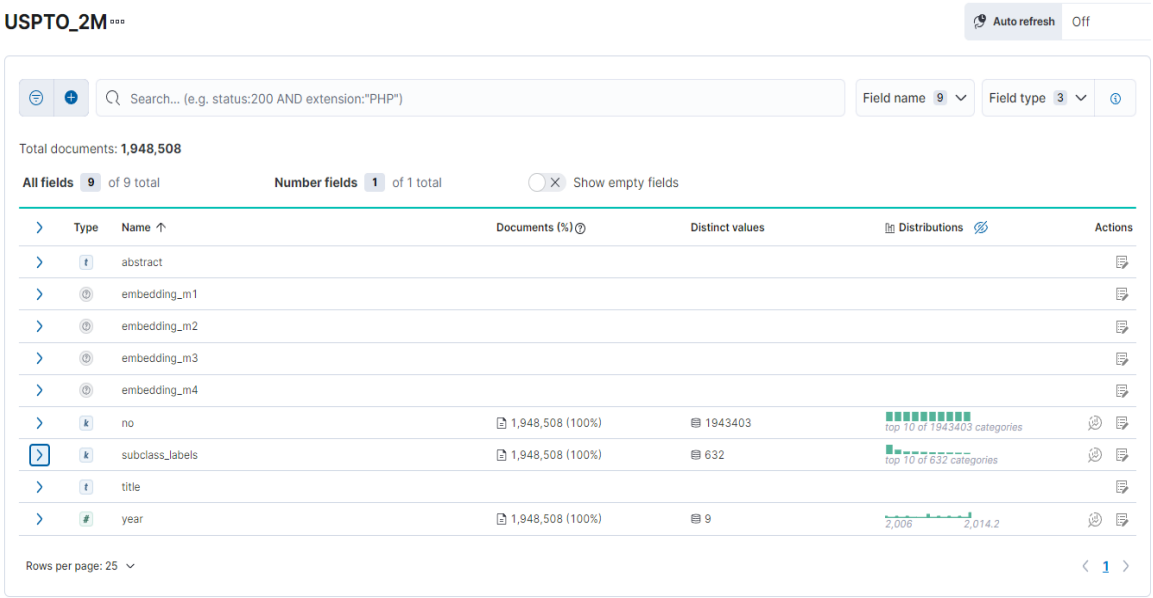
Figura 23 – Total de patentes indexadas entre 2006 e 2014



Fonte: elaborado pelo autor (2023)

Na Figura 24, tem-se o total de documentos indexados e a estrutura de armazenamento utilizada no Elasticsearch®. O conjunto de treinamento destaca-se por possuir 632 subclasses distintas em 9 anos de patentes armazenadas.

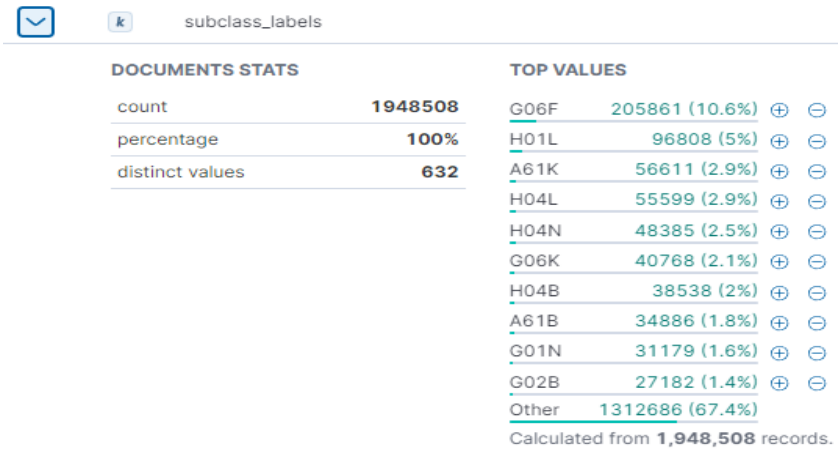
Figura 24 – Total de documentos indexados e estrutura de armazenamento



Fonte: elaborado pelo autor (2023)

Na Figura 25, é possível perceber que as patentes do conjunto de treinamento estão distribuídas em 623 subclasses distintas. Ademais, tem-se ainda 10 subclasses mais frequentes, sendo as subclasses G06F (10,60%), H01L (5,00%), A61K (2,90%), H04L (2,90%) e H04N (2,50%) as 5 mais frequentes. Dessa forma, nota-se que as subclasses não estão balanceadas, representando um desafio para qualquer estratégia de classificação.

Figura 25 – Informações sobre as subclasses mais frequentes



Fonte: elaborado pelo autor (2023)

Observa-se que as áreas tecnológicas avaliadas no período estão bem diversificadas – G06F (Computação; Processamento de dados) e H01L (Dispositivos semicondutores) são áreas tecnológicas com rápido desenvolvimento no período, impulsionando mais patentes. Já as patentes de subclasses A61K (Preparações para finalidades médicas), H04L (Transmissão de comunicações digitais) e H04N (Transmissão de imagens) são setores estabelecidos, com tecnologias e processos amplamente difundidos, o que limita o patenteamento de novidades e inovações incrementais. Ademais, percebe-se que o desequilíbrio no número de patentes entre subclasses reflete aspectos econômicos, legais e tecnológicos particulares a cada área. A análise da distribuição de patentes por subclasse fornece percepções importantes a respeito de tendências e inovações nas respectivas áreas tecnológicas.

3.4.3.5 Estratégias de ordenação

As estratégias de ordenação para elaborar o *ranking* das subclasses utilizando o valor de similaridade entre as patentes foram: a) soma das ocorrências (SO), ou seja, a frequência com que determinada subclasse ocorre no conjunto de documentos retornados; e b) soma dos

scores (SS), ou seja, para cada subclasse é computada a soma dos *scores* dos documentos aos quais a subclasse pertence.

Seja:

- D = {*d1*, *d2*, ..., *dn*} o conjunto de documentos retornados;
- C = {*c1*, *c2*, ..., *cm*} o conjunto de subclasses possíveis;
- *f*(*di*, *cj*) a frequência com que a subclasse *cj* ocorre no documento *di*;
- *s*(*di*) o *score* do documento *di*.

A equação para a estratégia Soma das Ocorrências (Equação 1) para a subclasse *cj* é:

$$SO(cj) = \sum f(di, cj) \quad (1)$$

Ou seja, somam-se as frequências da subclasse *cj* em todos os documentos retornados.

Já a equação para calcular a Soma dos *Scores* (SS) (Equação 2) para a subclasse *cj* é:

$$SS(cj) = \sum s(di).f(di, cj) \quad (2)$$

Ou seja, para cada documento, multiplica-se o seu *score* pela ocorrência da subclasse *cj*, acumulando-se esses valores. Assim, SO prioriza as subclasses mais frequentes no conjunto de documentos recuperados, enquanto SS prioriza subclasses que estejam associadas aos documentos recuperados de *score* mais elevado.

3.4.3.6 Similaridade vetorial

A similaridade vetorial permite medir o ângulo entre dois vetores, entendida a partir disso como uma similaridade entre dois vetores. Existem diversas equações que possibilitam aferir tal ângulo, mas geralmente se utiliza a distância do cosseno entre vetores, como apresentado na equação[24] a seguir:

[24] A equação foi adaptada de Shahid *et al*. (2020).

$$\cos(\theta) = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \cdot \sqrt{\sum_{i=1}^{n} B_i^2}} \quad (3)$$

Na equação acima, A e B são os vetores, *n* é o número de elementos dos vetores, A_i e B_i são os valores do *i*-ésimo elemento dos vetores, e θ é o ângulo entre os vetores. A similaridade do cosseno varia entre -1 e 1. Quanto mais próximo de 1, maior é a similaridade entre os vetores, e quanto mais próximo de -1, menor é a similaridade. Se a similaridade for 0, significa que os vetores são ortogonais ou perpendiculares, ou seja, não têm nenhuma relação.

3.4.3.7 Explicitação do conhecimento

A explicitação do conhecimento ocorre através da ordenação de subclasses de patentes (*ranking*) e, a partir disso, o grafo de conhecimento é gerado. O grafo de conhecimento estabelece um relacionamento entre tópicos e subclasses. Ao se prover um meio de visualizar os tópicos interconectados às subclasses mais relevantes sugeridas, obtém-se uma forma de explicitação que permite que examinadores tenham maior clareza do resultado quando determinada subclasse é apresentada pelo modelo proposto.

3.4.4 Demonstração

A atividade de demonstração da viabilidade do modelo proposto é executada em dois cenários de estudo, um mais geral com o conjunto total e outro mais específico com um conjunto reduzido de patentes. Em ambos os casos, os conjuntos de dados de patentes foram divididos em treinamento e teste. O conjunto de dados utilizado está descrito na seção 3.4.3.4.

3.4.5 Avaliação

Tradicionalmente, para a aferição de classificadores, são utilizadas métricas que avaliam de algumas maneiras acertos e erros para cada instância apresentada. Quando uma determinada instância é apresentada, tem-se uma expectativa, e no recebimento do resultado ocorre a análise para verificar a concordância, ou seja, se a resposta é correta ou errada. Métricas comuns utilizadas são acurácia (*accuracy*), precisão (*precision*), revocação (*recall*) e *f1-score*.

Essas métricas são definidas com base em uma matriz confusão[25] para problemas de classificação binária. O Quadro 12 apresenta uma matriz confusão e suas variáveis, onde *PV* representa os positivos verdadeiros e *NV* os negativos verdadeiros, *FP* corresponde aos falsos positivos e *FN* aos falsos negativos. Para cenários em que seja necessária a avaliação de mais classes, a matriz confusão pode ser adaptada expandindo a visão dicotômica tradicional de positivo e negativo.

Quadro 12 – Matriz confusão utilizada na avaliação da tarefa de classificação

		Classificação prevista	
		Positivo	Negativo
Classificação atual	Positivo	PV	FN
	Negativo	FP	NV

Fonte: elaborado pelo autor (2023)

As métricas derivadas a partir da matriz confusão são tradicionalmente aplicadas em cenários de classificação dicotômica e multiclasse, em que no conjunto de dados existem múltiplas classes, mas tanto a entrada quanto a saída possuem somente uma classe. Essas métricas ainda podem ser adaptadas para cenários multirrótulo (*multi-label*), conforme descrito em Mao, Tsang e Gao (2013).

Todavia, o cenário no qual esta tese se insere é entendido como multissaída (*multi-ouput*), assim como multientrada (*multi-input*). Ou seja, uma patente de entrada no conjunto de teste pode possuir uma ou mais subclasses, enquanto a resposta resultante é constituída de múltiplas subclasses ordenadas (*ranking*). Nesse sentido, métricas tradicionais poderiam ser aplicadas com alguma adaptação, mas os resultados não são capazes de refletir o desempenho de um modelo de classificação com foco em recomendação.

No tocante à métrica de avaliação utilizada no modelo, o ideal seria averiguar a correlação de *ranking*, tais como *Spearman's rank correlation coefficient* ou *Kendall rank correlation coefficient.* Todavia, as subclasses associadas a cada patente não indicam qualquer ordem, inviabilizando a aferição da correlação entre a ordem da patente de entrada e o *ranking* sugerido pelo modelo.

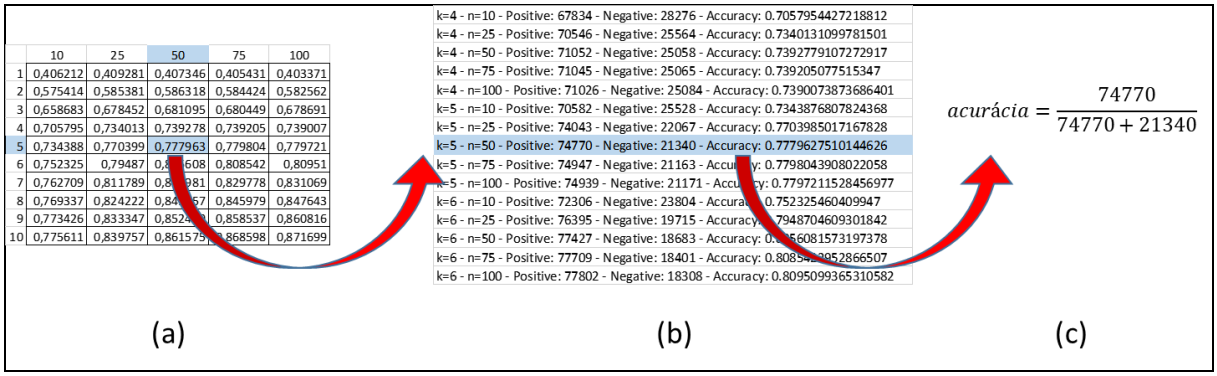
[25] Na literatura, o termo também é referenciado como matriz de confusão ou matriz de erro.

Com base nos aspectos acima mencionados, optou-se pela utilização de uma métrica de acurácia que leva em consideração o conjunto de subclasses recomendadas. Dessa forma, dada uma patente de entrada com 1 (uma) ou mais subclasses e uma recomendação (saída) com *k* subclasses, verifica-se se cada subclasse de entrada está no *ranking*. Em caso afirmativo, incrementa-se uma variável que representa o total de acertos da etapa de teste; do contrário, incrementa-se uma variável que representa o total de erros. A Equação 4 apresenta a métrica de acurácia utilizada na avaliação do modelo:

$$acurácia = \frac{TA}{TA + TE} \tag{4}$$

Na Equação 4, TA representa o total de acertos durante a etapa de teste, e TE representa o total de erros, também da etapa de teste. A acurácia é então calculada dividindo-se o total de acertos pelo número total de classificações (total de acertos + total de erros), como apresentado na Figura 26.

Figura 26 – Cálculo da acurácia



Fonte: elaborado pelo autor (2023)

Na Figura 26, tem-se um exemplo do cálculo da acurácia referente ao PTM *all-MiniLM-L6-v2*, considerando *k*=5 e *n*=50 (Figura 26a). Isso significa dizer que para cada patente do conjunto de teste serão recuperadas as 50 patentes mais similares (n), visando ao final da avaliação do conjunto de teste a recomendação das 5 subclasses mais relevantes (k=5). Na Figura 26b, tem-se o total de predições corretas (*Positive*), sendo 74.770 acertos e 21.340 erros (*Negative*), o que totaliza 96.110 predições. Ou seja, dessas 96.110 predições realizadas pelo modelo (visto que uma patente pode ter mais de uma subclasse), em 74.770 delas o modelo acertou a subclasse avaliada entre as 5 subclasses recomendadas. Já em 21.340 predições, considerando as subclasses avaliadas, nenhuma foi localizada entre as 5 subclasses recomendadas.

3.4.6 Comunicação

A comunicação das análises e a discussão dos resultados desta pesquisa serão realizadas através de publicação científica na forma de artigos em periódicos ou conferências, e após a conclusão da pesquisa na forma do documento final de tese. Também se pretende efetuar a solicitação de registro do protótipo desenvolvido para a avaliação do modelo, constituído basicamente de um módulo de mapeamento da base de conhecimento, um módulo de indexação e um modelo de consulta que determina as acurácias das estratégias de *ranking* implementadas. Adicionalmente, existem módulos para manipular o grafo de conhecimento, incluindo a extração das entidades, a construção do grafo de conhecimento geral e a constituição do grafo de conhecimento para determinada recomendação ordenada de subclasses.

3.5 SÍNTESE DA METODOLOGIA DE PESQUISA

Neste capítulo foram apresentados os procedimentos metodológicos adotados na tese. Quanto à natureza, a pesquisa se caracteriza como aplicada com cunho tecnológico, e quanto aos objetivos é classificada como exploratória. A metodologia utilizada para a construção do modelo foi a Design Science Research Methodology (DSRM).

Por fim, foram descritas as etapas de desenvolvimento da pesquisa de maneira geral (o detalhamento consta nos capítulos 4 e 5, onde cada elemento será utilizado na prática) tendo como base a DSRM. A síntese das atividades é apresentada no Quadro 13.

Quadro 13 – Síntese das atividades desenvolvidas na pesquisa

Atividades	**Descrição**
Identificar o problema e a motivação	Revisão integrativa da literatura.
Definir os objetivos para a solução	Revisão integrativa da literatura.
Projetar e desenvolver	Coleta de dados: patentes do conjunto de dados USPTO-2M envolvendo patentes dos anos de 2006 a 2015.
	Pré-processamento: tokenização e remoção de *stopwords*.
	Transformação dos dados: geração de vetores densos por meio da técnica de incorporação de documentos (*document embedding*).
	Treinamento: conjunto de dados separados em treinamento e teste. As patentes dos anos de 2006 a 2014 foram utilizadas para treinamento, e as de 2015 para o teste do modelo proposto.
	Explicitação do conhecimento: *ranking* de patentes e grafo do conhecimento.
Realizar a demonstração	Cenário de estudo: utilização do conjunto de patentes para a aplicação do modelo com foco na sugestão de subclasses de patentes de maneira ordenada (*ranking*) através de redes neurais profundas, bem como a explicitação do conhecimento por meio de grafos de conhecimento, visando subsidiar a escolha das subclasses mais adequadas para representar determinada patente.
Realizar a avaliação	É utilizada uma métrica de acurácia que analisa a(s) pertinência(s) da(s) subclasse(s) da patente de entrada no *ranking* de subclasses recomendado.
Realizar a comunicação	Apresentação dos resultados e considerações finais, assim como publicação de artigos e registro de software.

Fonte: elaborado pelo autor (2023)

4 MODELO PROPOSTO

Este capítulo apresenta, de forma detalhada, as etapas e o funcionamento do modelo proposto nesta tese. Ressalta-se a partir da revisão integrativa da literatura que não foram identificados trabalhos que considerassem a sugestão ordenada de subclasses de patentes, tampouco uma forma de explicitar o conhecimento envolto nas sugestões para auxiliar os especialistas (examinadores) no processo de análise e identificação das melhores subclasses. Ademais, o modelo também prevê a atualização da base de patentes, de modo que ele tenha um comportamento dinâmico ao longo do tempo.

Após a apresentação geral do modelo, é realizada uma instanciação, em que são apresentados os componentes técnicos e tecnológicos visando clarificar todas as etapas e suas interconexões.

4.1 APRESENTAÇÃO DO MODELO

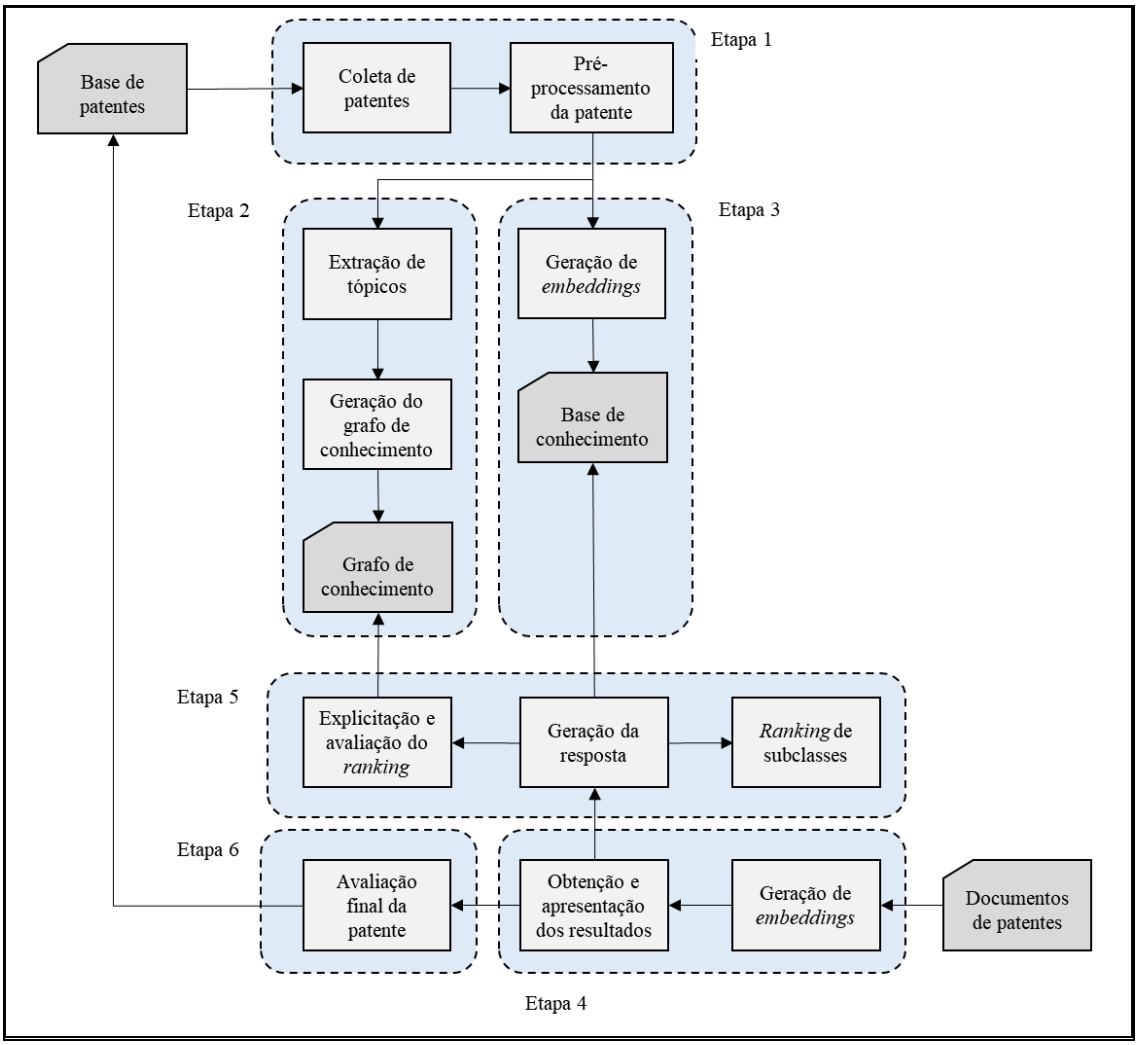
O modelo proposto compreende uma série de etapas, elaboradas com base na revisão integrativa da literatura e na fundamentação teórica, com a finalidade de responder à pergunta de pesquisa e atingir os objetivos geral e específicos. A seguir, serão apresentadas as etapas do modelo com uma breve descrição, conforme a Figura 27.

O objetivo final deste trabalho é propor um modelo voltado à recomendação de subclasses – mais especificamente subclasses de patentes – a partir de fontes de dados não estruturados na forma de texto (documentos de patentes) levando-se em conta aspectos de ordenação de subclasses (*ranking*) e explicitação de conhecimento. Ou seja, o *ranking* deve promover subsídios para o examinador de patentes, indicando quais são as subclasses com a maior relevância para serem atribuídas a determinada patente, bem como deve explicitar as relações que se formam entre os conceitos e as subclasses sugeridas, de maneira a auxiliar na tomada de decisão.

A Etapa 1 consiste na escolha de uma base de patentes disponível para realização de testes. A coleta de patentes pode ocorrer a partir de diversas fontes, entre elas a base de patentes disponibilizada pelo USPTO®. Após a coleta, descrita na seção 3.4.3.1, os dados passam por um conjunto de operações de pré-processamento, descritos na seção 3.4.3.2, sendo basicamente utilizada a remoção de eventuais pontuações e *stopwords*.

Na Etapa 2, são extraídos os tópicos mais relevantes que servirão de entrada para a geração do grafo de conhecimento, o qual possui como função essencial a explicitação do conhecimento envolto nas sugestões de subclasses a partir de determinada patente de interesse.

Figura 27 – Etapas do modelo proposto



Fonte: elaborado pelo autor (2023)

Já na Etapa 3 ocorre a geração dos *embeddings*, ou seja, os textos das patentes são transformados por diferentes PTMs com arquitetura *transformer*, sendo representados na forma de um vetor denso *n*-dimensional e armazenados em uma base de dados, designada base de conhecimento. Todavia, para que isso ocorra, primeiro é necessário realizar o mapeamento da estrutura do índice e, principalmente, a definição de como o *embedding* será armazenado. Após esse passo, ocorre a indexação, em que o documento e seu respectivo *embedding* são adicionados ao índice da base de conhecimento, permitindo que consultas sejam posteriormente realizadas.

A Etapa 4 é responsável pela avaliação dos documentos de patentes, os quais representam as patentes que ainda não possuem classificação. Levando em conta determinada demanda, uma patente que não tenha sido avaliada por um examinador passa por uma transformação, consistindo na geração do seu *embedding*. O *embedding* do documento da patente é então enviado para o processo de "Obtenção e apresentação dos resultados", sendo gerada uma consulta para obtenção do *ranking* de subclasses e o grafo de conhecimento correspondentes.

Na Etapa 5, a "Geração da resposta" tem por finalidade receber determinado *embedding* de documento de patente e avaliar a patente de acordo com a similaridade, ou seja, realiza-se uma consulta vetorial na base de conhecimento (que representa o conjunto de treinamento) para determinar os documentos (patentes) mais similares. Vale mencionar que, no fluxo de avaliação do modelo, os documentos de patentes utilizados nessa etapa constituem o conjunto de teste.

Após a consulta efetuada, invoca-se o processo de geração do "*Ranking* de subclasses" sob demanda, ou seja, o processo é ativado a partir da demanda de avaliação de algum novo documento de patente. O *ranking* fornece um conjunto ordenado de subclasses de patentes, apresentando a relação de subclasses da mais relevante para a menos relevante. Já a explicitação e a avaliação do *ranking* promovem suporte para a criação do grafo de conhecimento.

A partir do *ranking* de subclasses obtidas, ativa-se o processo de "Explicitação e avaliação do *ranking*", que consulta o grafo de conhecimento com as subclasses de patentes e como estas se interconectam através dos tópicos extraídos a partir das patentes com o intuito de facilitar o entendimento das sugestões ordenadas de subclasses. O processo de "Geração da resposta" devolve para o processo "Obtenção e apresentação dos resultados" o *ranking* de subclasses e o grafo de conhecimento.

Por fim, a Etapa 6 possui como objetivo a "Avaliação final da patente" a partir dos dados gerados na Etapa 4, isto é, com base no *ranking* e no grafo de conhecimento, tem-se um ferramental para auxiliar na tomada de decisão. Após a tomada de decisão pelo examinador da patente, ou seja, a escolha das subclasses mais adequadas, o resultado final composto pelo novo documento de patentes e suas subclasses é incorporado à base de patentes.

Em vista disso, esse contexto visa prover ferramental relevante aos examinadores, de modo a reduzir o tempo de avaliação de uma patente e aumentar a acurácia, ou seja, objetiva fornecer subsídios para decisões mais assertivas. Como ação final, determinada patente retorna para a base, agora com as subclasses escolhidas pelo examinador, o que impacta na atualização

do modelo de classificação e no grafo de conhecimento, permitindo, assim, a evolução das recomendações ao longo do tempo.

4.2 COMPOSIÇÃO DO MODELO

Nas subseções a seguir são apresentados mais detalhes para cada uma das etapas do modelo proposto.

4.2.1 Etapa 1: Coleta de dados e pré-processamento

Como mencionado, esta etapa é responsável pela coleta de patentes. As patentes podem ser coletadas a partir de diferentes fontes de dados abertos na *web* disponibilizados, por exemplo, pelo USPTO®. Detalhes são fornecidos na seção 3.4.3.1, que apresenta um conjunto de dados de referência em classificação de patentes elaborado por Li *et al.* (2018) com pouco mais de dois milhões de documentos de patentes de utilidade. Os documentos foram reunidos no período de 2006 a 2015, sendo compostos por subclasses, resumo, título e número da patente, conforme consta na seção 3.4.3.4.

Após a coleta de patentes, os dados passam por uma fase de pré-processamento e transformação, conforme descrito nas seções 3.4.3.2 e 3.4.3.3, respectivamente.

4.2.2 Etapa 2: Geração dos grafos de conhecimento

O grafo de conhecimento é obtido através de técnicas de NLP utilizadas na extração de tópicos e relacionamentos, em que os nós do grafo são representados por entidades (tópicos e subclasses), e as arestas descrevem o relacionamento entre essas entidades.

Após a análise de coocorrência, é construído o KG geral representando subclasses e seus tópicos associados. Apesar de existir um grafo geral, é possível extrair o KG de cada subclasse ou mesmo de algumas subclasses, de modo que este sirva de elemento fundamental na explicitação e num possível auxílio na compreensão do *ranking* de subclasses de patentes.

4.2.3 Etapa 3: Geração de *embeddings*

Nesta etapa, o conjunto de documentos de patentes é transformado em vetores densos (*embeddings*) por meio da utilização de um modelo de linguagem pré-treinado (PTM). Ou seja,

o texto que representa determinada patente passa por operações algébricas, de tal maneira que os vetores de patentes que possuem alta dimensionalidade sejam transformados em vetores densos, também chamados de *embeddings.* Os *embeddings* servem de suporte para a recomendação ordenada de subclasses, uma das tarefas centrais do modelo proposto.

4.2.4 Etapa 4: Avaliação, obtenção e apresentação dos resultados

Na Etapa 4, os documentos de patentes ainda sem classificação (código IPC) são apresentados ao modelo proposto. A "Geração de *embeddings*" consiste em gerar o *embedding* de determinado documento de patente, isto é, obtém-se a representação vetorial da patente de entrada utilizando-se um PTM (similar ao que ocorre na Etapa 3) com o intuito de comparar, com base na similaridade, os documentos mais aderentes armazenados na base de conhecimento.

O próximo passo é a "Obtenção e apresentação dos resultados", que consiste na utilização do *embedding*, gerado no passo anterior, a ser enviado à Etapa 5. Com o *embedding* gerado da patente de interesse é possível consultar e recuperar as patentes mais similares em relação àquelas que constam na base de conhecimento.

Como resultado, o examinador de patentes tem à sua disposição uma lista de subclasses ordenadas, na forma de um *ranking*, e um grafo de conhecimento composto pelas subclasses retornadas interconectadas com seus conceitos mais relevantes. Vale mencionar que os conceitos no KG se conectam entre si e com as subclasses. Com base nos resultados obtidos, o examinador possui elementos que podem auxiliá-lo na tomada de decisão, ou seja, na determinação da(s) subclasse(s) mais relevantes para a patente de interesse.

4.2.5 Etapa 5: Geração da resposta

A etapa de "Geração da Resposta" é uma das principais do modelo, pois realiza consultas na base de conhecimento utilizando buscas aproximadas que calculam a similaridade entre o vetor que representa determinada patente de entrada e os demais vetores de patentes disponíveis na base de conhecimento. O *embedding* do documento de patente, gerado na Etapa 4, serve de base para calcular a similaridade com os *embeddings* do conjunto de treinamento armazenados na base de conhecimento. Após a consulta, considerando o conjunto de patentes recuperadas, o "*Ranking* de subclasses" é produzido usando alguma estratégia de ordenação.

Neste trabalho, duas estratégias foram implementadas e são discutidas em detalhes na seção 3.4.3.5.

O *ranking* facilita a identificação, por parte dos examinadores, da relevância de cada subclasse associada a uma patente. Para tal, conforme mencionado na etapa anterior e considerando uma patente em particular, as estratégias de *ranking* fornecem uma lista de subclasses com suas respectivas relevâncias.

Já no processo de "Explicitação e avaliação do *ranking*", ocorre a consulta ao grafo de conhecimento levando em conta as subclasses que compõem o *ranking*, como descrito na seção 3.4.3.8. Por fim, o processo de "Geração da resposta" retorna ao processo "Obtenção e apresentação dos resultados" o *ranking* de subclasses e o grafo de conhecimento.

Considerando um fluxo normal de avaliação de patentes, o examinador recebe como resultado o *ranking* de subclasses com a respectiva relevância dessas subclasses. Ele possui, ainda, com base no conjunto de subclasses, acesso ao grafo de conhecimento. Ou seja, a partir de cada subclasse sugerida pela etapa anterior o examinador pode acessar o grafo de conhecimento, em que as subclasses mais relevantes sugeridas na etapa anterior são apresentadas em destaque, assim como os conceitos que as interconectam. Ademais, o grafo de conhecimento também destaca os tópicos que fazem parte da patente em análise. A possibilidade de investigação dos conceitos envoltos na análise por meio de um grafo de conhecimento objetiva tornar mais claro o resultado produzido pelo *ranking*.

4.2.6 Etapa 6: Avaliação final da patente

Por fim, com base no *ranking* de subclasses e no grafo de conhecimento, o examinador possui insumos para auxiliá-lo na tarefa de definir as subclasses mais adequadas para determinado documento de patente. Após a classificação do documento de patente, o modelo prevê a atualização da base de patentes com as subclasses selecionadas pelo examinador. Ou seja, considerando o conjunto de subclasses definidas como as mais representativas pelo examinador, estas são vinculadas a determinada patente na base de dados. Tal vinculação permite que a nova patente incorporada à base de dados seja insumo para novas recomendações visto que será enviada à base de conhecimento e ainda terá os principais conceitos extraídos para a atualização do grafo de conhecimento. Dessa forma, tem-se um ciclo de manutenção do conhecimento nos mais variados domínios envoltos no cenário de análise de patentes, assim como um contexto de aprendizado constante.

4.3 INSTANCIAÇÃO DO MODELO

O propósito desta seção é demonstrar, na forma de um exemplo, a instanciação do modelo proposto com vistas a clarificar como se comporta. Dessa maneira, pretende-se demonstrar os componentes do modelo e a ligação entre eles, assim como as técnicas e tecnologias envolvidas. O conteúdo textual utilizado para a instanciação das etapas do modelo proposto é apresentado no Quadro 14, sendo a patente US08001811 considerada como exemplo.

Quadro 14 – Exemplificação do conteúdo de uma patente

	Número da patente US08001811
Título	*washing machine having water softening device*
Resumo	*washing machine water softening device improves solubility detergent water softening performance concurrently washing machine includes tub water supply device supplying water tub detergent supply device supplying detergent tub water softening device softening water water softening device disposed water supplied water supply device mixed detergent supplied detergent supply device water mixed detergent supplied water softening device washing machine water softening device*
Subclasse	D06F

Fonte: elaborado pelo autor a partir da patente US08001811 (2023)

No Quadro 15, tem-se o conjunto de técnicas, tecnologias/ferramentas, a representação vetorial e a estratégia de ordenação aplicadas em cada uma das etapas do modelo proposto.

Quadro 15 – Elementos utilizados na instanciação do modelo

		Tecnologia/Ferramenta					Representação Vetorial		Estratégias de Ordenação	
Etapa	Subetapa	Python	Elasticsearch	Gephi	spaCy	DBpedia	Sentence-BERT	Tok2Vec	Soma das ocorrências	Soma dos *scores*
Etapa 1	Base de patentes	X								
	Coleta de patentes	X								
	Pré-Processamento da patente	X								
Etapa 2	Extração de tópicos	X			X	X				
	Geração do grafo de conhecimento			X						
	Grafo de conhecimento			X						
Etapa 3	Geração de *embeddings*	X					X	X		
	Base de conhecimento	X	X							
Etapa 4	Documentos de patentes	X								
	Avaliação dos documentos de patentes	X					X	X		
	Obtenção e apresentação dos resultados	X								
Etapa 5	Geração da resposta	X	X							
	Ranking de subclasses	X							X	X
	Explicitação e avaliação do *ranking*	X		X						
Etapa 6	Avaliação final da patente	X								

Fonte: elaborado pelo autor (2023)

4.3.1 Etapa 1: Coleta de dados e pré-processamento

Como visto anteriormente, a base de dados utilizada para a instanciação e avaliação do modelo foi a USPTO-2M. De modo geral, esse conjunto de dados passou pelo processo de limpeza, descrito nas seções 3.4.3.1 e 3.4.3.2. De agora em diante, a referência aos dados será realizada por meio da palavra *corpus*, conjunto de dados formado pelas patentes extraídas do conjunto USPTO-2M.

Na etapa de pré-processamento, utilizou-se a linguagem Python® e a biblioteca NLTK® (Natural Language Toolkit) para efetuar a remoção de *stopwords*, que consiste em remover as palavras irrelevantes ou que não carregam um significado relevante para a análise do texto, como pronomes, preposições e artigos. Para realizar testes no modelo proposto, utilizou-se um conjunto de dados com o texto completo, conforme disponibilizado por Li *et al*. (2018), e outro conjunto de dados aplicando-se a remoção de *stopwords*.

4.3.2 Etapa 2: Geração dos grafos de conhecimento

Após a limpeza dos dados, as entidades da patente US08001811 foram extraídas e nomeadas utilizando o spaCy® – o resultado é apresentado no Quadro 16. O NER identifica as entidades nomeadas, vinculando a elas conceitos relevantes, assim como realiza o mapeamento para uma base de dados de destino. Nesse caso, a base de dados[26] escolhida para efetuar esse mapeamento e produzir resultados mais consistentes foi a DBpedia®.

Tal estratégia permite lidar com a desambiguação de entidades e o mapeamento de conceitos, enriquecendo o grafo de conhecimento por meio de informações sobre os conceitos obtidas a partir da base de conhecimento DBpedia®.

Quadro 16 – Representação dos conceitos extraídos utilizando NER

Conceitos		
Descrição	URI	Ocorrências
detergent	http://dbpedia.org/resource/Detergent	6
solubility	http://dbpedia.org/resource/Solubility	1
washing machine	http://dbpedia.org/resource/Washing_machine	4
water softening	http://dbpedia.org/resource/Water_softening	7

Fonte: elaborado pelo autor (2023)

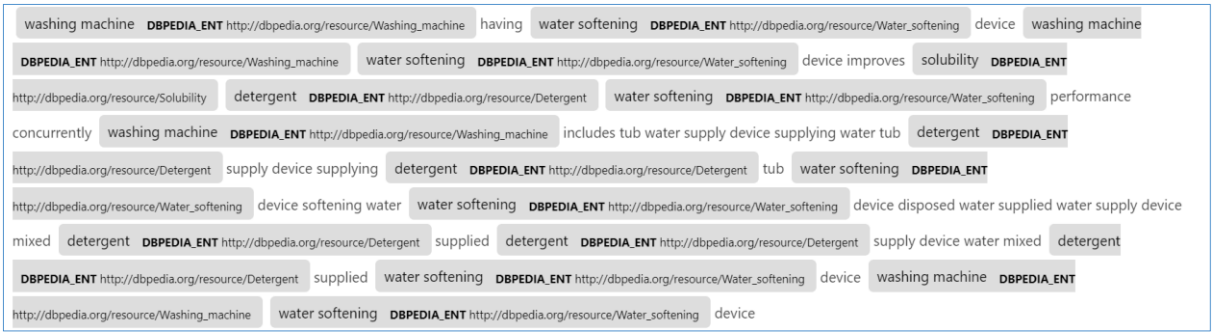
Foram identificadas 4 entidades, sendo *detergent* com 6 ocorrências, *solubility* com 1 ocorrência, *washing machine* com 4 ocorrências e *water softening* com 7 ocorrências. Os

[26] Aqui entendida como base de conhecimento em razão da sua amplitude semântica ao organizar diversas características de determinado conceito, bem como integrar essas características com dados não estruturados.

conceitos desses termos podem ser acessados por meio do URI (Uniform Resource Identifier) disponível através da base de conhecimento da DBpedia®.

Na Figura 28, são apresentadas as entidades nomeadas e seus rótulos destacados no texto da patente.

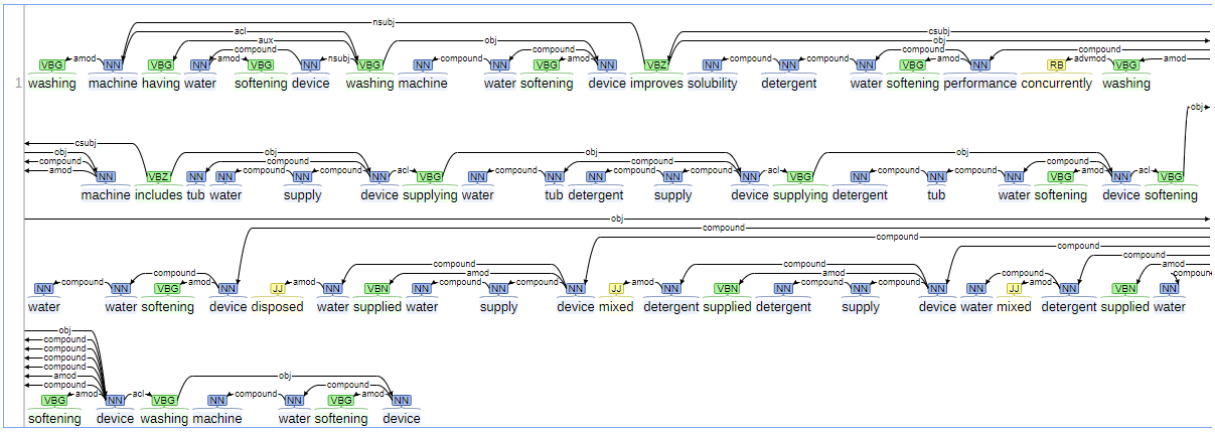
Figura 28 – Entidades nomeadas e rótulos no texto da patente



Fonte: elaborado pelo autor (2023)

O restante das palavras da lista de *tokens* foi classificado conforme suas subclasses gramaticais. As dependências sintáticas entre os *tokens* e a subclasse gramatical são apresentadas na Figura 29. Foi utilizada a ferramenta CoreNLP®[27] para uma melhor visualização das dependências em que a representação sintática do texto mostra as relações entre os *tokens*. O resultado dessa etapa é um vetor de tópicos que servirá de entrada para a geração do grafo de conhecimento.

Figura 29 – Dependências sintáticas



Fonte: obtido a partir da patente de exemplo utilizando a ferramenta CoreNLP®

Para a construção do grafo de conhecimento, os elementos mais importantes são os nós e as arestas, sendo os nós representados pelas entidades (tópicos e subclasses) e as arestas pelos

[27] Disponível em: https://stanfordnlp.github.io/CoreNLP. Acesso em: 28 maio. 2023. O CoreNLP® recebe o texto bruto, executa uma série de anotadores em NLP no texto e devolve um conjunto de anotações.

relacionamentos que conectam tópicos com subclasses. Para gerar o grafo de conhecimento é preciso seguir algumas etapas, de modo a converter os dados não estruturados em um grafo. Nesse sentido, a partir do texto da patente cada entidade extraída é relacionada às subclasses que constam na patente, atribuindo a quantidade de vezes (frequência) que os tópicos aparecem no texto, conforme o Quadro 16. Assim, à medida que cada patente é analisada, o grafo evolui. O Quadro 17 apresenta um exemplo a partir da patente US08001811 com a subclasse D06F e considerando outra patente que possui outra subclasse, aqui identificada como H04J.

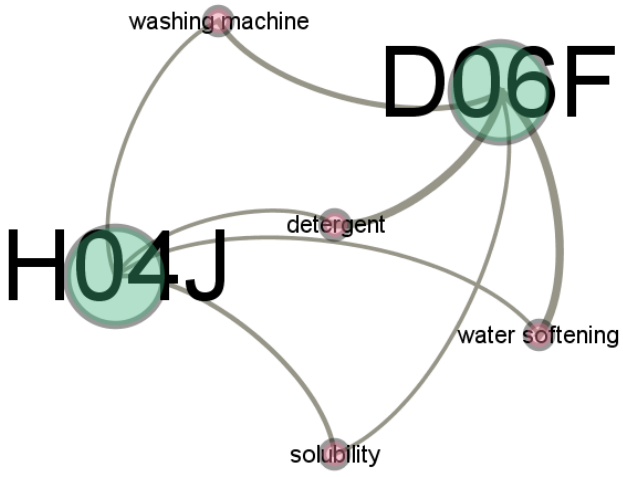
Quadro 17 – Frequência dos tópicos extraídos associados às subclasses

Tópico	**Frequências para a subclasse D06F**	**Frequências para a subclasse H04J**
detergent	6	1
solubility	1	2
washing machine	4	1
water softening	7	1

Fonte: elaborado pelo autor (2023)

O Quadro 17 indica determinado tópico associado à(s) subclasse(s) com as respectivas frequências. Cada tópico pode estar associado a *n* subclasses. Uma vez que os dados estejam integrados, um grafo de conhecimento pode ser gerado, conforme ilustra a Figura 30.

Figura 30 – Grafo de conhecimento com tópicos associados às subclasses



Fonte: elaborado pelo autor (2023)

Pretende-se com o KG promover a visualização geral dos tópicos do domínio e também os tópicos vinculados à(s) subclasse(s), como descrito na seção 2.3.1 a partir da recomendação das subclasses aos examinadores. Com o propósito de visualizar essa etapa da

instanciação, a ferramenta Gephi® foi utilizada para gerar a representação dos grafos de conhecimento. De modo geral, a representação foi criada de forma que seja possível observar as subclasses de patentes, os tópicos dessas subclasses e os relacionamentos entre tópicos e subclasses.

4.3.3 Etapa 3: Geração de *embeddings* e armazenamento na base de conhecimento

Nesta etapa, realiza-se o mapeamento dos dados, a indexação e a geração dos *embeddings*. Durante o mapeamento, define-se o tipo de cada campo da patente, por exemplo, *"no"* para o número da patente, *"subclass_labels"* para a subclasse da patente, *"text"* para título e resumo da patente. Também são especificados os campos para o armazenamento dos vetores densos, por exemplo, *"embedding_m1"*, *"embedding_m2"*, *"embedding_m3"* e *"embedding_m4"*, sendo voltados à realização das pesquisas e dos cálculos de similaridade. Cada um dos quatro campos de vetores densos está associado a determinado PTM. Esses vetores densos foram utilizados na instanciação do modelo proposto.

A transformação dos dados textuais, do título e do resumo em *embeddings* e posterior indexação permite que a busca e a recuperação de informações relevantes em documentos de patentes sejam mais eficientes. Para tal, algumas bibliotecas são necessárias, entre elas a PyTorch®, disponível na linguagem Python®. O PyTorch® é uma biblioteca de cálculo de tensores e aceleração de Graphics Processing Units (GPU) utilizada para aplicações de ML e DL. Fornece uma maneira conveniente de definir e de treinar muitas arquiteturas de DL, sendo criada com o objetivo de possibilitar uma experimentação rápida.

Como mencionado, a geração dos *embeddings* se dá a partir da parte textual da patente disponível no conjunto de dados da USPTO-2M, ou seja, o título e o resumo. Com base nesse conteúdo e na utilização de PTMs ocorre a transformação em vetores densos. O Quadro 18 apresenta os PTMs utilizados para instanciar o modelo proposto.

Quadro 18 – Modelos pré-treinados utilizados na instanciação

Modelo	**Tamanho**	**Dimensão**	**Base**	**Tipo**
all-mpnet-base-v2	420 MB	768	Sentence-BERT	Contextual
all-distilroberta-v1	290 MB	768	Sentence-BERT	Contextual
all-MiniLM-L6-v2	80 MB	384	Sentence-BERT	Contextual
en_core_web_lg	382 MB	300	Tok2Vec	Estático

Fonte: elaborado pelo autor (2023)

Os modelos que possuem *"all-*"* no prefixo foram treinados com mais de um bilhão de pares, sendo projetados para uso geral e fazendo parte do projeto SBERT® (Sentence-

BERT). Apesar do nome, o SBERT® permite a geração de *embeddings* para textos curtos, podendo, dessa forma, ser também classificado como *document embedding*. Esses modelos foram amplamente avaliados pela qualidade na geração de vetores densos de sentenças (Performance Sentence Embeddings) e para consultas de pesquisa em vetores densos (Performance Semantic Search). O modelo *all-mpnet-base-v2* proporciona melhor qualidade, já o modelo *all-MiniLM-L6-v2* é menor e mais rápido (em torno de cinco vezes), e ainda assim atinge resultados equivalentes quando comparado a PTMs maiores (Reimers; Gurevych, 2019).

De modo geral, o SBERT aprimora o BERT usando uma rede siamesa para codificar cada sentença em uma matriz de alta dimensão. Essa rede siamesa consiste em várias camadas que transformam a entrada em uma representação vetorial densa. O modelo é treinado para que as representações de sentenças semanticamente semelhantes sejam mapeadas para vetores próximos uns dos outros, enquanto as representações de sentenças semanticamente diferentes sejam mapeadas para vetores distantes (Reimers; Gurevych, 2019).

Já o PTM *en_core_web_lg* faz parte do *pipeline* do spaCy®, treinado em um extenso conjunto de dados para algumas línguas. O modelo é otimizado para CPU e contém componentes como "*tok2vec*", "*tagger*", "*parser*", "*senter*", "*ner*", "*attribute_ruler*" e "*lemmatizer*". O componente Tok2Vec é responsável por gerar vetores de palavras, que são representações de alta dimensionalidade das palavras de determinada sentença. A partir de operações sobre o conjunto de vetores das palavras da sentença, esse componente possibilita a incorporação de documentos, ou seja, a geração do *embedding* de documentos (spaCy, 2023).

No Quadro 19, tem-se um exemplo da representação vetorial utilizando o PTM *all-MiniLM-L6-v2* de 384 dimensões, apresentados os 15 primeiros e os 15 últimos valores do *embedding* produzido a partir da patente US08001811.

Quadro 19 – Representação vetorial utilizando *all-MiniLM-L6-v2*

-8.824071	-3.203836	1.073805	-5.999068	7.791073
-6.458390	-7.330572	-1.67299	-1.379745	-5.94362
3.0823933	1.4056314	1.049880	3.1645372	3.813410
…				
-1.138514	2.7878334	1.847830	8.4616661	-2.196646
-3.808791	8.3671316	-4.03556	-6.322803	6.4507275
8.5419245	1.6600670	4.428521	5.5819209	-4.776708

Fonte: elaborado pelo autor (2023)

Por fim, após a geração dos *embeddings,* estes são armazenados em um banco de dados NoSQL que provê funcionalidades para armazenar vetores densos, também chamado de banco de dados vetorial. Para este trabalho, utilizou-se o Elasticsearch®, um banco de dados gratuito,

distribuído e orientado a documentos, e que armazena dados em formato JSON. O Elasticsearch® permite o mapeamento de diferentes tipos de dados, tais como números, datas, textos e vetores densos, possibilitando combinar buscas estruturadas e não estruturadas e, no caso dos vetores densos, buscas aproximadas, característica esta que promove uma semântica mais apurada aos resultados. De modo geral, representa a base de conhecimento do modelo proposto, proporcionando a recuperação aproximada de patentes, tarefa necessária para a geração de recomendações de subclasses.

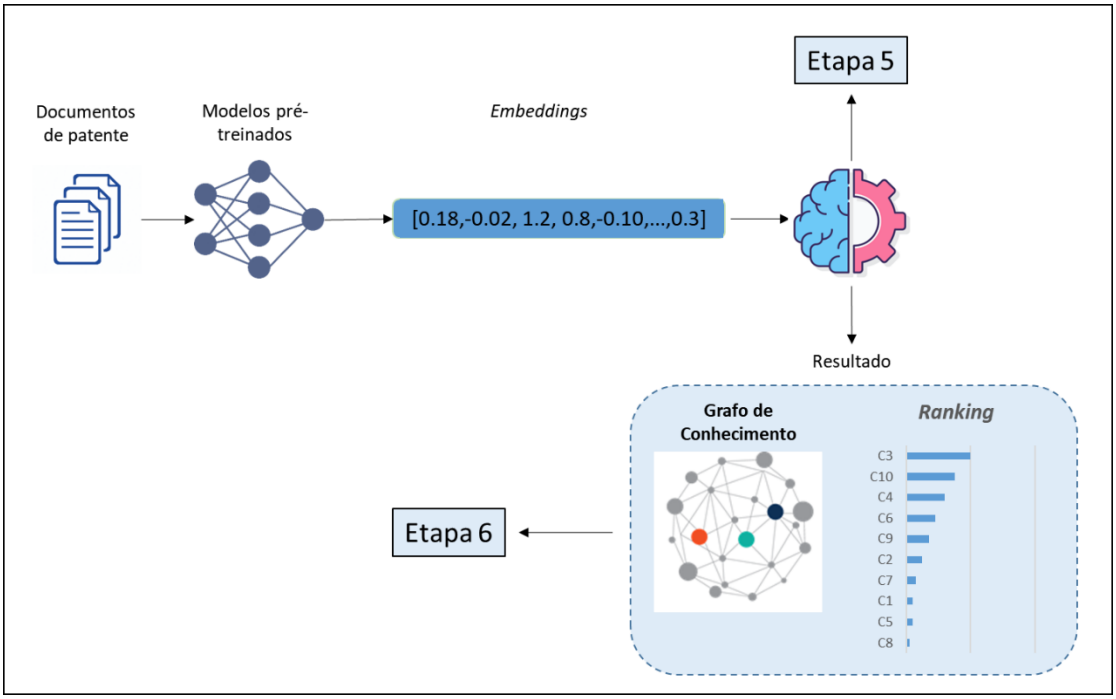
4.3.4 Etapa 4: Avaliação, obtenção e apresentação dos resultados

Etapa que tem por finalidade receber os documentos de patentes que fazem parte do conjunto de teste e avaliar essas patentes de acordo com a Etapa 5, ou seja, faz-se uma consulta na base de conhecimento para determinar os documentos mais semelhantes. Assim como os demais componentes do modelo proposto, esta etapa foi desenvolvida com a linguagem Python®, tendo como objetivo recomendar o *ranking* de subclasses e o respectivo grafo de conhecimento para subsidiar examinadores em sua tomada de decisão.

Na Figura 31, tem-se a representação da Etapa 4, em que os documentos de patente compostos por título e resumo são transformados em *embeddings.* Cada patente é transformada de acordo com o PTM correspondente, conforme descrito no Quadro 18. Os *embeddings* são então enviados para a Etapa 5, de modo que seja realizada uma busca vetorial aproximada utilizando estratégias de ApNN. O modelo, através dos *embeddings* gerados na Etapa 4, permite identificar as patentes mais similares com base nas representações armazenadas na base de conhecimento por meio de um cálculo de similaridade. A partir do *ranking* de subclasses, também é obtido o KG, levando em conta as subclasses recomendadas.

Como resultado dessa etapa, tem-se uma lista ordenada de subclasses (*ranking*) e um grafo de conhecimento que explicita informações sobre as subclasses relevantes. Essas informações fornecem subsídios para o examinador tomar decisões relacionadas às subclasses que devem ser atribuídas a determinada patente.

Figura 31 – Representação da Etapa 4



Fonte: elaborado pelo autor (2023)

4.3.5 Etapa 5: Geração da resposta

A Etapa 5 é executada tendo como entrada um *embedding* obtido através da Etapa 4. O objetivo consiste na realização de uma consulta na base de conhecimento para determinar os documentos mais semelhantes.

A geração de resposta é ativada através de uma *query*, solicitada na Etapa 4, para então o modelo realizar uma consulta no conjunto de treinamento, ou seja, na base de conhecimento, para determinar os documentos mais semelhantes. O módulo de consulta foi desenvolvido em Python® visando realizar operações de consulta e análise de dados no ambiente Elasticsearch®.

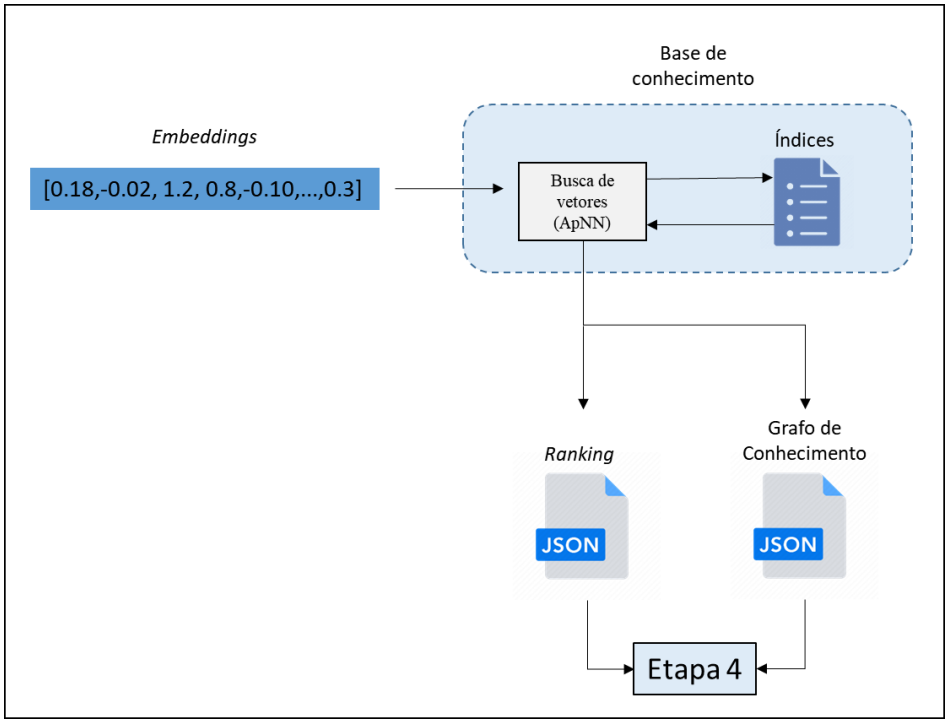
A Figura 32 exibe o funcionamento da Etapa 5, em que é realizada uma busca vetorial, em que os *embeddings* são utilizados na consulta das patentes indexadas na base de conhecimento. A consulta é realizada levando em conta o conceito de busca aproximada por meio da estratégia ApNN, que consiste em consultar os dados indexados para encontrar os *k*-vizinhos mais próximos (KNN), isto é, as patentes mais próximas considerando uma patente de entrada representada pelo seu *embedding* enviado pela Etapa 4.

Com o resultado da busca por similaridade é possível elaborar o *ranking* de patentes, representado por uma lista ordenada com as subclasses de patentes obtidas com base na lista de documentos relevantes e em determinada estratégia de ordenação. Cabe mencionar que o

ranking é produzido por demanda, ou seja, é gerado sempre que existir a necessidade de classificação de uma nova patente.

A recomendação ordenada de subclasses é realizada através de duas estratégias de ordenação (seção 3.4.3.5): 1) a soma das ocorrências (SO); e 2) a soma dos *scores* (SS).

Figura 32 – Representação da Etapa 5



Fonte: elaborado pelo autor (2023)

O Quadro 20 apresenta a relação de subclasses recomendadas e suas respectivas relevâncias. Nesse caso, considerou-se o modelo *all-mpnet-base-v2* e a estratégia de ordenação SS da patente nº US08472379, onde consta a indicação da subclasse e a sua frequência com sua respectiva relevância. O resultado leva em consideração um $k = 5$, isto é, foram recomendadas as 5 subclasses mais frequentes, e $n = 50$, ou seja, os 50 documentos mais similares. Para este k e n, as 5 subclasses foram mencionadas 71 vezes nos 50 documentos. Ressalta-se que a lista de subclasses é maior, atingindo 78 ocorrências nos 50 documentos, todavia somente as 5 subclasses mais relevantes foram apresentadas. Assim, a subclasse H04W tem uma frequência de 34, ou seja, foi mencionada em 34 dos 50 documentos em um total de 78 referências, totalizando 43,59%.

Quadro 20 – Relação da relevância das subclasses

Subclasse	H04W	H04B	H04L	H04J	H04Q
Frequência	34	13	11	10	3
Relevância (%)	43,59	16,67	14,10	12,82	3,85

Fonte: elaborado pelo autor (2023)

Além da recomendação do *ranking* de subclasses, a avaliação dessa recomendação pode ser realizada através de um grafo de conhecimento (KG). Cabe mencionar que, no exemplo do Quadro 20, o número de subclasses foi limitado a 5, embora a relação seja maior. Dessa forma, considerando uma interface de usuário, um examinador poderia requisitar a apresentação do KG com todas as subclasses do *ranking*, todavia estando em destaque somente as *k* subclasses selecionadas por ele como as mais relevantes.

De modo geral, o KG permite ao examinador explorar o resultado do *ranking* visualizando as interconexões entre as subclasses através dos tópicos. Além disso, o KG também enfatiza os tópicos que estão presentes na patente em análise. A possibilidade de investigação com o uso desse tipo de representação tem o objetivo de tornar mais claro o resultado produzido. Destaca-se que, para esta etapa, não se desenvolveu um módulo de representação na forma de grafo/rede, sendo utilizada simplesmente a ferramenta Gephi®. Todavia, o grafo de conhecimento geral e os individuais, conforme a demanda, foram salvos em formato GraphML, sendo possível a visualização na ferramenta Gephi®.

4.3.6 Etapa 6: Avaliação final das patentes

A Etapa 6 corresponde à avaliação final da patente, em que são utilizados os dados gerados nas etapas anteriores para auxiliar o examinador na tomada de decisão. Nessa fase, o objetivo é fornecer subsídios para uma classificação mais precisa e embasada nas subclasses mais relevantes à patente em questão.

O examinador tem acesso ao *ranking* de subclasses previamente gerado, no qual apresenta, para determinada patente de interesse, um conjunto ordenado (*ranking*) de subclasses por sua relevância, servindo de subsídio à seleção das subclasses mais apropriadas pelos examinadores. Além disso, o KG é empregado como uma ferramenta visual para auxiliar na compreensão das relações entre as subclasses sugeridas e os tópicos que as compõem. Ou seja, através da análise do grafo de conhecimento vislumbra-se ser possível examinar de forma ampla a estrutura do conhecimento envolto na recomendação das subclasses.

Com base nessas informações, o examinador pode realizar uma análise criteriosa para determinar as subclasses mais adequadas para a patente em questão. Essas subclasses selecionadas refletem o conhecimento especializado do examinador, resultando em uma classificação mais precisa e relevante. Após a escolha das subclasses, estas são atribuídas à patente e incorporadas à base de patentes, promovendo a atualização da base de conhecimento, a extração de possíveis novos tópicos e a consequente incorporação ao grafo de conhecimento. Essa atualização contínua permite que o modelo evolua ao longo do tempo, tendo a expectativa de aprimorar as sugestões e promover resultados mais precisos para os examinadores.

Portanto, a Etapa 6 desempenha um papel fundamental ao fornecer suporte técnico para a tomada de decisão do examinador, combinando o conhecimento especializado com as informações geradas pelo sistema, a fim de realizar uma classificação mais embasada e precisa das subclasses de patentes.

4.4 CONSIDERAÇÕES FINAIS

Este capítulo apresentou o modelo proposto, detalhando o funcionamento de todas as suas etapas. Demonstrou-se a coleta e a transformação dos dados para viabilizá-lo, passando pela geração do grafo de conhecimento, geração dos *embeddings* e composição da base de conhecimento, com o intuito de recomendar um conjunto ordenado (*ranking*) de subclasses, assim como explicitar as relações entre subclasses e tópicos, de forma a permitir que o examinador tenha uma visão mais ampla do resultado ofertado.

As etapas do modelo proposto foram apresentadas considerando elementos mais técnicos, a fim de clarificar o seu funcionamento por meio de uma instanciação. Os componentes do modelo, a conexão entre eles e as técnicas/tecnologias envolvidas também foram detalhados. A apresentação e a discussão dos resultados obtidos na instanciação do modelo serão apresentadas no próximo capítulo.

5 ANÁLISE E DISCUSSÃO DOS RESULTADOS

Este capítulo exibe os resultados obtidos por intermédio da instanciação do modelo proposto utilizando o conjunto de dados apresentado na seção 3.4.3.1 e a avaliação abordada na seção 3.4.5. Nesse sentido, a análise dos resultados objetiva discutir a instanciação do modelo por meio de PTMs e *embeddings* de documento como técnica de representação vetorial, assim como uma base de conhecimento para viabilizar consultas aos *embeddings* com o intuito de recomendar subclasses de maneira ordenada. Ademais, concentra-se na utilização de grafos de conhecimento para auxiliar no entendimento das subclasses sugeridas e escolha das melhores subclasses por examinadores, assim como na atualização das avaliações efetuadas impactando nos resultados do modelo proposto.

5.1 AMBIENTE DE AVALIAÇÃO DO MODELO PROPOSTO

Visando clarificar o ambiente computacional utilizado nesta tese, no que tange ao conjunto de experimentos descritos na seção 5.2 deste capítulo, os seguintes componentes foram considerados:

- Equipamento: para o presente trabalho, utilizou-se uma *workstation* Lenovo P330 Intel Xeon E-2174G (04 core/08 threads de 3.8 - 4.7 GHz) com 16 GB de RAM DDR4, 1 TB HD SATA, 512 GB HD SSD m.2 e Placa Quadro de 2 GB, DDR5;
- Algoritmos: foram avaliados 4 modelos pré-treinados (PTMs) baseados em arquiteturas de redes neurais do tipo *transformers*, sendo os seguintes: *all-MiniLM-L6-v2*, *all-mpnet-base-v2*, *all-distilroberta-v1* e *en_core_web_lg*. Também foram testadas as redes neurais de aprendizado profundo MLP, CNN e LSTM, que serviram de base para a primeira versão do modelo proposto e, na sequência, para a comparação com o modelo final;
- Configurações: cada PTM foi testado com *embeddings* de dimensões variadas, utilizando as estratégias de Soma das Ocorrências (SO) e Soma dos Scores (SS) para *ranking*. Variaram-se os parâmetros *n* (patentes retornadas) e *k* (subclasses sugeridas);
- Métricas: a métrica utilizada para avaliar o desempenho dos modelos foi a acurácia, calculada dividindo-se o total de acertos pelo total de acertos acrescido do total de erros;

- Conjuntos de dados: foram utilizadas patentes de 2006 a 2014 (aproximadamente 2 milhões) para o treinamento e patentes de 2015 (aproximadamente 50 mil) para o teste. Algumas patentes, como a US08472379, foram avaliadas individualmente para exemplificar e clarificar o comportamento do modelo proposto. A Figura 33 representa a estrutura de determinada patente armazenada e indexada na base de conhecimento utilizando o banco de dados Elasticsearch®. A Figura 34 apresenta a distribuição dos dados de treinamento e teste usados no cenário geral em que somente MPTs foram utilizados e no cenário em que MPTs foram comparados com redes neurais de aprendizado profundo (MLP, CNN e LSTM).

Figura 33 – Estrutura utilizada para armazenamento e indexação de patentes

Nome	Tipo	Pesquisável	Agregável	Descrição
_id	_id	•		Identificador único da patente
_index	_index	•	•	Índice do Elasticsearch onde o documento está armazenado
_score				Pontuação de relevância da patente para uma determinada consulta
_source	_source			Conteúdo completo da patente
abstract	text	•		Resumo textual do conteúdo do documento de patente.
embedding_m1	dense_vector	•		Vetor denso de *embedding* do texto da patente gerado pelo modelo MiniLM-L6-v2
embedding_m2	dense_vector	•		Vetor denso de *embedding* do texto da patente gerado pelo modelo mpnet-base-v2
embedding_m3	dense_vector	•		Vetor denso de *embedding* do texto da patente gerado pelo modelo distilroberta-v1
embedding_m4	dense_vector	•		Vetor denso de *embedding* do texto da patente gerado pelo modelo distilroberta-v1
no	keyword	•	•	Número identificador único do documento de patente
subclass_labels	keyword	•	•	Lista de *keywords/tags* que categorizam a patente (seção, classe e subclasse)
title	text	•		Título da patente
year	integer	•	•	Ano de publicação/concessão da patente

Fonte: elaborado pelo autor (2023)

Figura 34 – Distribuição dos dados nos dois cenários de avaliação

	Modelo com MPTs		Modelo com Redes Neurais	
Conjunto de dados USPTO-2M	**Treinamento**	**Teste**	**Treinamento**	**Teste**
Coleta de dados	2006-2014	2015	2012, 2013 e 2014	2012, 2013 e 2014
Pré-processamento	texto em minúsculo; remoção de pontuação; remoção de *stopwords*	texto em minúsculo; remoção de pontuação; remoção de *stopwords*	texto em minúsculo; remoção de pontuação; remoção de *stopwords*	texto em minúsculo; remoção de pontuação; remoção de *stopwords*
Transformação dos dados	*Embeddings*	*Embeddings*	Tokenização	Tokenização
Treinamento	1.948.508	49.900	40.000	10.000
Estratégias de ordenação	Soma das ocorências e Soma dos *scores*	Soma das ocorências e Soma dos *scores*	Soma das probabilidades	Soma das probabilidades
Similaridade vetorial	Cosseno	Cosseno		
Explicitação do conhecimento	*Ranking* e KG	*Ranking* e KG	*Ranking* e KG	*Ranking* e KG

Fonte: elaborado pelo autor (2023)

O uso da Placa Quadro (placa de vídeo) auxiliou na indexação através do acelerador de processamento gráfico (GPU) com o uso do CUDA (Compute Unified Device Architeture) criado pela NVIDIA®. A biblioteca NVIDIA CUDA® Deep Neural Network (cuDNN) também

foi utilizada para a aceleração da GPU para redes neurais profundas. Ressalta-se que a instalação do CUDA e cuDNN depende da configuração do equipamento que será utilizado para o processamento.

5.2 APRESENTAÇÃO DOS RESULTADOS

A avaliação do modelo proposto está centrada na utilização das patentes e nas transformações que servem de entrada para redes neurais do tipo *transformer* (utilizada através de PTMs). Para a etapa de treinamento voltado à avaliação do modelo proposto, os PTMs descritos no Quadro 18 foram utilizados para geração e indexação dos *embeddings* na base de conhecimento para as patentes dos anos de 2006 a 2014. A Tabela 1 apresenta os resultados da indexação, exibindo ainda o total de patentes indexadas e o tempo gasto para a indexação de aproximadamente 2 (dois) milhões de patentes.

A coluna “Patentes” diz respeito ao número de patentes armazenadas em um arquivo em determinado ano. A coluna “Indexadas” se refere ao número exato de patentes indexadas. A coluna “Diferença” é a subtração entre a coluna “Patentes” e a coluna “Indexadas”. Essa diferença ocorre devido à falta, em algumas patentes para o referido arquivo de determinado ano, de conteúdo para o campo *“subclass_labels”*, descrito na seção 3.4.3.1. No total, 1.739 (mil setecentas e trinta e nove) patentes não foram indexadas para o treinamento.

Tabela 1 – Indexação das patentes

Ano	Patentes	Indexadas	Diferença	Tempo em minutos	Média/minuto
2006	175.499	174.942	557	335	522,21
2007	158.896	158.522	374	303	523,17
2008	159.669	159.459	210	303	526,27
2009	168.923	168.801	122	320	527,50
2010	221.761	221.615	146	419	528,91
2011	226.718	226.591	127	404	560,87
2012	255.293	255.181	112	482	529,42
2013	280.154	280.065	89	530	528,42
2014	303.334	303.332	2	567	534,98
Total	**1.950.247**	**1.948.508**	**1.739**	**3.663**	531,94

Fonte: elaborado pelo autor (2023)

O número de patentes indexadas foi de 1.948.508 (um milhão, novecentas e quarenta e oito mil, quinhentas e oito) em um tempo de 3.663 (três mil, seiscentos e sessenta e três) minutos, aproximadamente 61 horas de execução para indexar o conjunto de dados de patentes.

As patentes indexadas representam a base de conhecimento do modelo proposto que suporta a recomendação ordenada de subclasses. Já a coluna “Média/minuto” indica a quantidade média de patentes indexadas por minuto. Como média final, tem-se que aproximadamente 531 (quinhentas e trinta e uma) patentes foram indexadas por minuto. As colunas “Tempo em minutos” e “Média/minuto” dependem do equipamento utilizado para a indexação.

A partir da indexação do conjunto de dados, foi possível realizar a instanciação e a execução do modelo proposto. Nesse sentido, iniciou-se a fase de testes do modelo, que contou com um total de 49.900 (quarenta e nove mil e novecentas) patentes oriundas do conjunto de dados de patentes do ano de 2015. O conjunto de dados possui, ao todo, 1.998.408 (um milhão, novecentas e noventa e oito mil, quatrocentas e oito) patentes de treinamento e de teste.

As análises realizadas foram obtidas a partir de 800 execuções (instanciações) do modelo, compostas pelas combinações de diferentes parâmetros. Esse valor resulta da multiplicação dos PTMs utilizados na geração dos *embeddings*, com as estratégias de ordenação, a variação do parâmetro *k* (número máximo de subclasses recomendadas, variando de 1 até 10), número de documentos recuperados em cada iteração de *k*, parâmetro *n* (valores: 10, 25, 50, 75, 100) e diferentes formas de pré-processamento. Desse modo, o cálculo é realizado multiplicando-se 4 PTMs, 2 estratégias de *ranking* (SO e SS), 10 posições de *ranking* (*k*), 5 quantidades de documentos recuperadas *n* em cada iteração de *k* e 2 formas de pré-processamento, sendo 4 x 2 x 10 x 5 x 2, o que totaliza 800 instanciações.

No Quadro 21, tem-se o detalhamento das instâncias utilizadas no cenário de estudo, divididas em parâmetros e elementos utilizados. Apresenta informações sobre o conjunto de dados USPTO-2M e os modelos pré-treinados usados para gerar os *embeddings* dos dados para os modelos *all-mpnet-base-v2*, *all-distilroberta-v1*, *all-MiniLM-L6-v2* e *en_core_web_lg*.

Quadro 21 – Detalhamento das instâncias utilizadas no cenário de estudo

Parâmetros	**Elementos**
Conjunto de dados	USPTO-2M
Modelo utilizado para geração do *embedding*	a) *all-mpnet-base-v2* b) *all-distilroberta-v1* c) *all-MiniLM-L6-v2* d) *en_core_web_lg*
Conteúdo	Título e resumo
Pré-processamento	- Remoção de pontuação (a) - Transformação do conteúdo para minúsculo (b) - Retirada de *stopwords* (c) Pré-processamento 1 (a + b) Pré-processamento 2 (a + b + c)
Estratégia de ordenação	a) Soma das ocorrências b) Soma dos *scores*
n	10; 25; 50; 75; 100
k	1; 2; 3; 4; 5; 6; 7; 8; 9; 10

Fonte: elaborado pelo autor (2023)

O conteúdo textual dos dados consiste em título e resumo, que passaram por pré-processamentos envolvendo a remoção de pontuação e a transformação das *strings* para letras minúsculas e retirada de *stopwords*. A estratégia de ordenação dos dados baseou-se na soma das ocorrências (SO) e na soma dos *scores* (SS), descritas na seção 3.4.3.5.

O valor de *n* representa o número de documentos recuperados em uma determinada consulta (patente de entrada) para o cálculo do *ranking* considerando uma estratégia em particular. Já o valor de *k* representa o número de subclasses relevantes, ou seja, as *k* subclasses mais relevantes a serem recomendadas para uma determinada patente de entrada.

A quantidade de patentes por subclasse é pouco relevante para o modelo, visto que bastariam algumas patentes (considerando o menor *n* utilizado, que foi 10) mencionando determinada subclasse para que fosse definida como a mais similar. Isso ocorre porque uma patente de teste é transformada em um vetor denso, e depois são localizadas as patentes mais similares que permitem a composição do *ranking* que será recomendado. Isso mostra a capacidade de generalização do modelo, que, mesmo com poucos dados para uma determinada subclasse, é capaz de atingir uma acurácia interessante.

Ademais, quando o modelo é comparado às redes neurais, os resultados são muito superiores, o que demonstra, de maneira inequívoca, que o modelo lida adequadamente com subclasses com poucas patentes.

Considerando as configurações do Quadro 21, foram realizadas as avaliações com o conjunto de teste, sendo os resultados apresentados separadamente na próxima seção.

5.2.1 Avaliação dos PTMs

A base de conhecimento do modelo proposto contém a representação vetorial dos PTMs. Essa estrutura serve de suporte para a realização de consultas (*querys*), e o modelo proposto utiliza um conjunto de teste objetivando aferir as acurácias nas diferentes instanciações realizadas. Nesta seção, serão detalhados os resultados gerais e específicos levando-se em conta as diferentes combinações de instanciações para gerar as acurácias e, dessa forma, avaliar o modelo.

A acurácia é realizada conforme descrito na seção 3.4.5. Com base em determinada patente, a métrica é calculada verificando se suas subclasses pertencem ao conjunto de subclasses retornadas em cada recomendação, sendo as patentes recuperadas da base de conhecimento.

De modo geral, ainda que existam outros parâmetros, os dois mais relevantes considerando determinado PTM, estratégia de ordenação e tipo de pré-processamento, são o *k* e o *n*. Assim, para cada valor de *k* (quantidade de subclasses recomendadas) e *n* (quantidade de documentos avaliados para a determinação do *ranking*), caso determinada subclasse da patente de entrada seja encontrada na lista de patentes recomendada, considera-se como uma classificação correta (verdadeiro positivo), do contrário, como uma classificação incorreta (falso positivo).

A Tabela 2 apresenta um detalhamento das acurácias obtidas utilizando os diferentes PTMs para a estratégia SO. Os valores máximo, mínimo, média, mediana e desvio-padrão das acurácias calculadas com a variação de *k* e *n* são apresentados. Os resultados obtidos nas análises a seguir consideram o texto completo (título e resumo) sem a retirada de *stopwords*.

Tabela 2 – Indicadores estatísticos de acurácia para os diferentes PTMs com a estratégia de *ranking* SO

Acurácia (%)	**Modelo**			
	all-MiniLM-L6-v2	***all-mpnet-base-v2***	***all-distilroberta-v1***	***en_core_web_lg***
Mínima	40,43	40,63	40,16	25,50
Máxima	87,05	87,34	86,50	67,81
Média	72,68	72,98	72,22	52,58
Mediana	77,27	77,45	76,70	56,27
Desvio-padrão	13,48	13,52	13,42	12,17

Fonte: elaborado pelo autor (2023)

Os modelos *all-MiniLM-L6-v2*, *all-mpnet-base-v2* e *all-distilroberta-v1* apresentaram acurácias semelhantes mesmo com dimensões vetoriais diferentes, utilizando transformadores

de sentenças que levam em consideração o contexto das palavras componentes da sentença. Os modelos *all-mpnet-base-v2* e *all-distilroberta-v1* possuem dimensões vetoriais de 768 dimensões, enquanto o modelo *all-MiniLM-L6-v2* possui 386 dimensões.

O modelo *en_core_web_lg* apresentou uma acurácia bem abaixo dos demais modelos. Como esse modelo trabalha com *embeddings* estáticos e 300 dimensões, a incorporação de frases é construída calculando-se a média das incorporações de palavras.

A Tabela 3 apresenta as acurácias máxima, mínima, média, mediana e desvio-padrão para a estratégia de ordenação (*ranking*) utilizando a soma dos *scores* (SS) sem a retirada de *stopwords*.

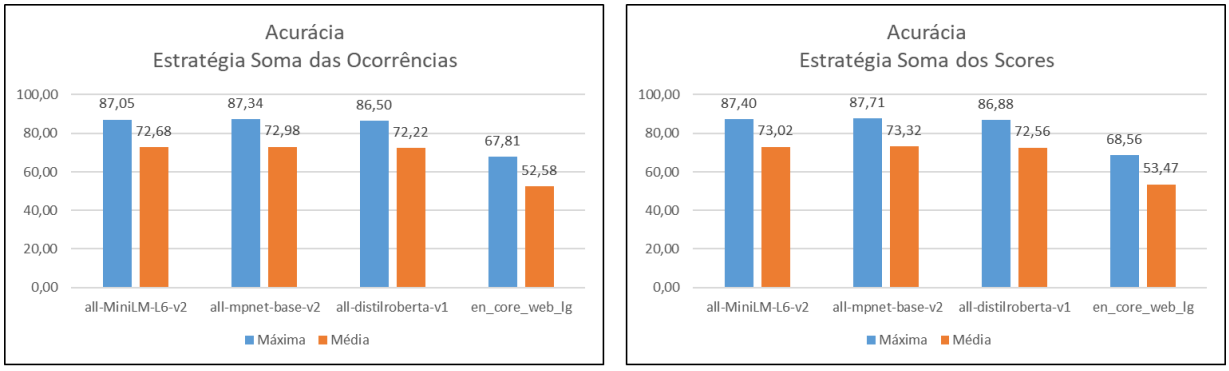
Tabela 3 – Indicadores estatísticos de acurácia para os diferentes PTMs com a estratégia de *ranking* SS

Acurácia (%)	Modelo			
	all-MiniLM-L6-v2	all-mpnet-base-v2	all-distilroberta-v1	en_core_web_lg
Mínima	40,49	40,70	40,21	26,07
Máxima	87,40	87,71	86,88	68,56
Média	73,02	73,32	72,56	53,47
Mediana	77,47	77,70	76,91	57,01
Desvio-padrão	13,58	13,62	13,52	12,35

Fonte: elaborado pelo autor (2023)

Os resultados das acurácias dos modelos são levemente superiores com o emprego da estratégia de ordenação SS. O Gráfico 1 exibe a comparação das acurácias entre as duas estratégias de ordenação utilizadas. O Gráfico 1(a) apresenta a acurácia dos modelos usando a estratégia de SO, e o Gráfico 1(b) a estratégia SS, sendo que a cor azul representa a acurácia máxima, e a cor laranja representa a acurácia média.

Gráfico 1 – Comparação das acurácias para as estratégias de *ranking*, SO e SS



(a) (b)

Fonte: elaborado pelo autor (2023)

Na Tabela 4 são demonstrados os resultados dos testes realizados para os quatro PTMs contendo as duas estratégias de ordenação para cada modelo. Como mencionado anteriormente, as acurácias são estabelecidas variando o número de documentos recuperados para a recomendação ordenada de subclasses (*n*) e a quantidade de subclasses recomendadas (*k*).

Os resultados demonstram que os modelos *all-MiniLM-L6-v2*, *all-mpnet-base-v2* e *all-distilroberta-v1* apresentam resultados superiores quando comparados com o modelo *en_core_web_lg*. Percebe-se ainda que os resultados para os modelos *all-MiniLM-L6-v2*, *all-mpnet-base-v2* e *all-distilroberta-v1* atingem certa estabilidade nas acurácias para *n*=50 e *k*=6, nas duas estratégias de ordenação, estando em torno de 80%. Valores maiores de *n* possuem pouco impacto no aumento da acurácia. Por outro lado, o aumento do *k* promove melhores acurácias, visto que são maiores as chances de alguma subclasse de uma patente em particular durante a etapa de teste estar presente na lista de recomendações.

A partir dos resultados, percebe-se que a estratégia de ordenação SS obteve um desempenho um pouco superior em relação à estratégia SO. Quanto maior o valor de *n* e *k*, maior a acurácia do modelo. Sendo assim, à medida que mais subclasses são consideradas, ou seja, que se aumenta a ordem do *ranking*, aumenta-se também a acurácia. Isso promove mais chances de localizar determinada subclasse da patente na lista ordenada de subclasses. Todavia, recomendar muitas subclasses pode não ser adequado para suportar a tomada de decisão de um examinador.

Tabela 4 – Acurácias dos testes gerais para os PTMs utilizados com diferentes configurações de *n* e *k* nas estratégias de *ranking* SO e SS

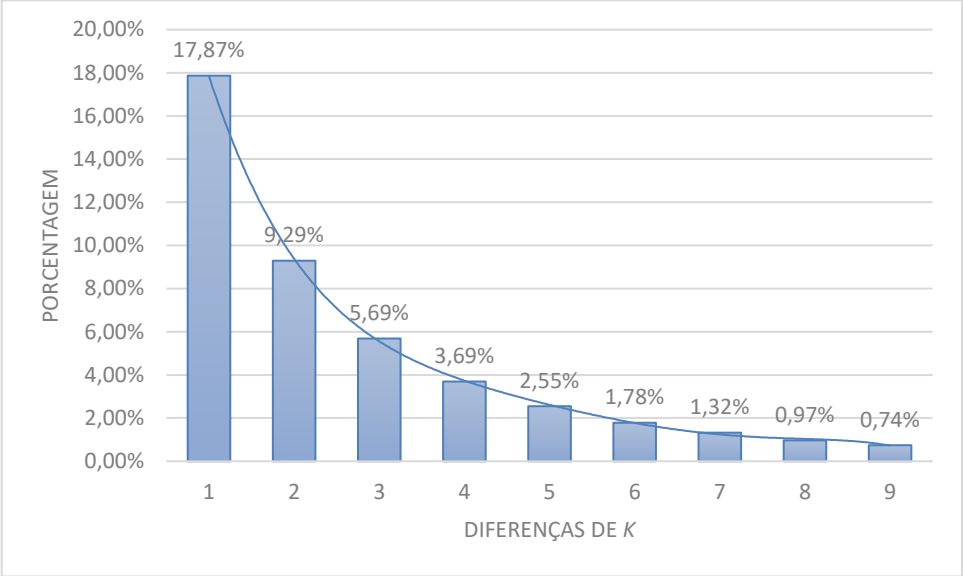
Modelo	Estratégia de ordenação	*n*	*k* =1	*k* =2	*k* =3	*k* =4	*k* =5	*k* =6	*k* =7	*k* =8	*k* =9	*k* =10
all-MiniLM-L6-v2	SO	10	40,57	57,22	65,58	70,25	73,11	75,03	76,28	77,03	77,51	77,83
		25	40,99	58,52	67,61	73,17	76,73	79,07	80,83	82,04	82,97	83,67
		50	40,78	58,67	68,14	73,99	77,77	80,47	82,46	83,97	85,08	85,95
		75	40,64	58,60	68,16	74,07	77,95	80,76	82,87	84,49	85,74	86,75
		100	40,43	58,43	68,00	73,97	78,03	80,91	83,04	84,72	85,98	87,05
	SS	10	40,71	57,65	66,10	70,89	73,76	75,49	76,63	77,28	77,66	77,91
		25	41,08	58,76	68,00	73,65	77,20	79,68	81,38	82,60	83,55	84,25
		50	40,85	58,83	68,41	74,35	78,20	80,94	82,89	84,34	85,49	86,41
		75	40,71	58,74	68,44	74,34	78,26	81,14	83,27	84,87	86,13	87,15
		100	40,49	58,54	68,19	74,19	78,22	81,21	83,41	85,02	86,37	87,40
all-mpnet-base-v2	SO	10	40,80	57,56	65,82	70,60	73,54	75,39	76,52	77,24	77,67	77,94
		25	41,11	58,87	68,01	73,60	77,12	79,54	81,23	82,47	83,37	84,01
		50	40,99	58,98	68,43	74,26	78,11	80,82	82,67	84,17	85,34	86,24
		75	40,77	58,83	68,36	74,38	78,35	81,15	83,21	84,77	86,01	87,01
		100	40,63	58,68	68,30	74,36	78,35	81,27	83,39	85,02	86,29	87,34
	SS	10	40,94	57,98	66,45	71,18	73,99	75,79	76,82	77,45	77,79	77,99
		25	41,22	59,12	68,37	74,02	77,62	80,04	81,77	82,93	83,80	84,47
		50	41,06	59,20	68,70	74,58	78,50	81,25	83,17	84,71	85,86	86,73
		75	40,84	58,98	68,56	74,59	78,65	81,54	83,62	85,26	86,44	87,40
		100	40,70	58,81	68,47	74,61	78,69	81,60	83,75	85,41	86,71	87,71
all-distilroberta-v1	SO	10	40,35	57,00	65,16	69,79	72,72	74,51	75,72	76,48	76,92	77,18
		25	40,63	58,15	67,11	72,74	76,23	78,65	80,38	81,59	82,52	83,20
		50	40,53	58,31	67,61	73,40	77,32	80,00	81,97	83,49	84,63	85,52
		75	40,35	58,19	67,66	73,56	77,54	80,27	82,35	83,97	85,21	86,19
		100	40,16	57,89	67,51	73,56	77,56	80,42	82,56	84,15	85,42	86,50
	SS	10	40,50	57,42	65,75	70,50	73,30	74,96	76,04	76,69	77,07	77,28
		25	40,70	58,36	67,53	73,16	76,75	79,21	80,88	82,10	82,98	83,62
		50	40,62	58,48	68,01	73,80	77,67	80,43	82,49	84,01	85,14	86,04
		75	40,39	58,31	67,92	73,85	77,84	80,63	82,76	84,40	85,66	86,63
		100	40,21	58,03	67,71	73,80	77,82	80,72	82,87	84,52	85,80	86,88
en_core_web_lg	SO	10	25,50	36,83	43,26	47,49	50,86	53,43	55,51	57,16	58,49	59,38
		25	26,46	38,70	45,85	50,95	54,73	57,55	59,69	61,39	62,69	63,85
		50	26,34	39,27	46,87	52,06	55,96	59,05	61,48	63,43	65,18	66,68
		75	26,13	39,36	47,08	52,55	56,57	59,69	62,15	64,19	65,90	67,40
		100	25,90	39,31	47,00	52,46	56,62	59,80	62,40	64,44	66,22	67,81
	SS	10	26,31	38,08	44,91	49,46	52,65	55,01	56,78	58,13	59,12	59,80
		25	26,86	39,37	46,79	51,97	55,84	58,77	61,09	62,97	64,46	65,68
		50	26,61	39,82	47,45	52,83	56,82	59,93	62,46	64,55	66,29	67,77
		75	26,36	39,73	47,54	53,09	57,21	60,39	62,94	65,04	66,82	68,39
		100	26,07	39,60	47,41	53,01	57,20	60,45	63,00	65,18	66,95	68,56

Fonte: elaborado pelo autor (2023)

Para uma visão inicial da importância de se recomendar um número que possa ser caracterizado como adequado de subclasses, o Gráfico 2 apresenta a curva das diferenças das acurácias variando o *k* entre 1 e 10 para *n*=50 e a estratégia de ordenação SS. O primeiro valor

refere-se à diferença do valor *k*=2 – *k*=1, o segundo valor *k*=3 – *k*-2, e assim por diante. Dessa forma, percebe-se certa estabilidade para um *k* entre 5 e 6.

Gráfico 2 – Média das diferenças percentuais das acurácias para o modelo *all-mpnet-base-v2*



Fonte: elaborado pelo autor (2023)

Já para o modelo *en_core_web_lg*, ainda que os resultados para as duas estratégias de ordenação sejam inferiores aos demais PTMs, o comportamento geral na variação do *n* é similar. Por outro lado, mesmo que o aumento do valor de *k* impacte positivamente nos resultados, o desempenho desse modelo é inferior.

Os testes iniciais foram realizados considerando o texto completo sem a remoção de *stopwords*. Com o intuito de verificar o comportamento do modelo proposto diante da remoção de *stopwords*, treinou-se o modelo novamente, removendo-se as *stopwords* do título e do resumo da patente. Ou seja, as quase 2 (duas) milhões de patentes foram transformadas e indexadas novamente removendo-se as *stopwords*. O resultado final da base de conhecimento com a remoção de *stopwords* é apresentado na Tabela 5, sendo possível perceber, de maneira geral, uma pequena melhora em todos os PTMs.

Tabela 5 – Acurácias dos testes gerais, com remoção de *stopwords*, para os PTMs utilizados com diferentes configurações de *n* e *k* nas estratégias de *ranking* SO e SS

Modelo	Estratégia de ordenação	*n*	*k* =1	*k* =2	*k* =3	*k* =4	*k* =5	*k* =6	*k* =7	*k* =8	*k* =9	*k* =10
all-MiniLM-L6-v2	SO	10	40,37	57,06	65,22	69,94	72,85	74,74	75,95	76,72	77,22	77,50
		25	40,82	58,24	67,33	72,92	76,47	78,86	80,53	81,80	82,71	83,45
		50	40,68	58,42	67,79	73,61	77,40	80,13	82,12	83,64	84,77	85,70
		75	40,48	58,32	67,77	73,67	77,67	80,50	82,64	84,20	85,42	86,38
		100	40,27	58,14	67,67	73,68	77,75	80,70	82,76	84,43	85,70	86,73
	SS	10	40,62	57,54	65,87	70,58	73,44	75,23	76,27	76,93	77,34	77,56
		25	40,93	58,54	67,85	73,40	77,04	79,49	81,18	82,42	83,33	83,98
		50	40,73	58,63	68,11	73,93	77,80	80,56	82,60	84,11	85,24	86,16
		75	40,54	58,44	68,04	73,92	77,98	80,85	82,98	84,60	85,85	86,86
		100	40,34	58,26	67,87	73,90	77,97	80,95	83,11	84,76	86,08	87,17
all-mpnet-base-v2	SO	10	40,74	57,45	65,70	70,56	73,42	75,19	76,27	77,03	77,47	77,75
		25	41,03	58,67	67,81	73,36	76,96	79,45	81,08	82,29	83,20	83,87
		50	40,96	58,96	68,37	74,25	78,16	80,87	82,80	84,30	85,39	86,25
		75	40,77	58,77	68,41	74,35	78,47	81,26	83,40	84,96	86,14	87,10
		100	40,52	58,60	68,25	74,40	78,47	81,35	83,48	85,12	86,41	87,46
	SS	10	40,93	57,94	66,36	71,10	73,93	75,61	76,64	77,24	77,61	77,80
		25	41,17	58,95	68,21	73,80	77,49	79,92	81,61	82,77	83,66	84,28
		50	41,02	59,09	68,69	74,54	78,52	81,26	83,25	84,73	85,84	86,71
		75	40,83	58,93	68,57	74,62	78,72	81,59	83,72	85,32	86,54	87,47
		100	40,58	58,69	68,43	74,62	78,73	81,66	83,78	85,47	86,76	87,77
all-distilroberta-v1	SO	10	40,09	56,52	64,64	69,50	72,40	74,30	75,47	76,23	76,73	76,99
		25	40,36	57,73	66,75	72,28	75,92	78,37	80,11	81,44	82,40	83,13
		50	40,13	57,90	67,21	73,12	77,06	79,82	81,78	83,27	84,42	85,33
		75	39,94	57,78	67,21	73,19	77,22	80,05	82,09	83,77	85,07	86,07
		100	39,76	57,53	67,02	73,14	77,23	80,15	82,31	83,89	85,25	86,29
	SS	10	40,24	56,93	65,22	70,03	72,91	74,67	75,81	76,46	76,85	77,06
		25	40,46	58,02	67,18	72,86	76,52	79,04	80,77	82,00	82,82	83,51
		50	40,22	58,09	67,54	73,47	77,46	80,23	82,26	83,77	84,96	85,84
		75	40,00	57,94	67,48	73,48	77,48	80,45	82,58	84,20	85,49	86,54
		100	39,80	57,65	67,22	73,38	77,48	80,48	82,63	84,30	85,68	86,74
en_core_web_lg	SO	10	31,07	44,25	51,54	56,12	59,47	61,92	63,84	65,29	66,27	66,93
		25	31,54	45,77	53,85	59,29	63,07	65,84	67,97	69,51	70,76	71,79
		50	31,44	46,16	54,58	60,26	64,23	67,23	69,50	71,42	72,94	74,24
		75	31,22	46,18	54,58	60,34	64,42	67,54	69,98	71,92	73,54	74,86
		100	31,00	45,95	54,52	60,35	64,51	67,63	70,11	72,12	73,77	75,15
	SS	10	31,57	45,23	52,93	57,79	61,02	63,26	64,84	65,95	66,66	67,17
		25	31,88	46,39	54,59	60,25	64,10	66,96	69,06	70,87	72,15	73,20
		50	31,64	46,55	55,02	60,86	64,88	67,95	70,30	72,22	73,82	75,18
		75	31,37	46,45	54,98	60,83	64,98	68,17	70,69	72,64	74,28	75,65
		100	31,10	46,16	54,81	60,72	64,99	68,15	70,68	72,73	74,39	75,83

Fonte: elaborado pelo autor (2023)

Nos modelos contextuais a alteração da acurácia, para mais ou para menos, considerando-se o texto completo sem e com a retirada de *stopwords*, não demonstra ser relevante. Já no caso do modelo estático *en_core_web_lg*, percebe-se uma melhora significativa. Para exemplificar, apresenta-se no Quadro 22 a acurácia obtida comparando-se o texto completo (título e resumo da patente) sem e com a remoção de *stopword* para n=100 e k=10.

Quadro 22 – Comparativo com a remoção ou não de *stopwords* para o modelo *en_core_web_lg*

Conteúdo	**Estratégia de ordenação**	**Acurácia**
Texto completo (sem remoção de *stopwords*)	SO	67,81
	SS	68,56
Texto completo (com remoção de *stopwords)*	SO	75,15
	SS	75,83

Fonte: elaborado pelo autor (2023)

Com a remoção de *stopwords*, o modelo *en_core_web_lg* obteve resultados bem mais significativos, independentemente da estratégia utilizada. A melhora na acurácia se deve à característica não contextual do modelo, que não considera as dependências entre palavras. De maneira geral, a incorporação é realizada através da média de todas as palavras diferentes. Com isso, muitas palavras (por exemplo, "*is*", "*the*", "*to*", "*of*", entre outras) estão semanticamente na mesma região no espaço n-dimensional. Essas palavras, ao serem incorporadas a determinada sentença, reduzem a sua representatividade semântica em um contexto específico.

As próximas seções mostram uma avaliação mais detalhada de cada um dos PTMs, sendo *all-MiniLM-L6-v2*, *all-mpnet-base-v2, all-distilroberta-v1* e *en_core_web_lg.*

5.2.1.1 Modelo all-MiniLM-L6-v2

A seguir, apresentam-se os resultados para o modelo *all-MiniLM-L6-v2* exibidos na Tabela 6. Entre os modelos contextuais este é o que permite a geração de *embeddings* com a menor dimensionalidade. Apesar disso, esses modelos mostram-se adequados em relação aos de maior dimensão, produzindo resultados equivalentes.

Tabela 6 – Acurácias para o modelo *all-MiniLM-L6-v2*

Modelo	Estratégia de ordenação	*n*	*k* =1	*k* =2	*k* =3	*k* =4	*k* =5	*k* =6	*k* =7	*k* =8	*k* =9	*k* =10
all-MiniLM-L6-v2	SO	10	40,57	57,22	65,58	70,25	73,11	75,03	76,28	77,03	77,51	77,83
		25	40,99	58,52	67,61	73,17	76,73	79,07	80,83	82,04	82,97	83,67
		50	40,78	58,67	68,14	73,99	77,77	80,47	82,46	83,97	85,08	85,95
		75	40,64	58,60	68,16	74,07	77,95	80,76	82,87	84,49	85,74	86,75
		100	40,43	58,43	68,00	73,97	78,03	80,91	83,04	84,72	85,98	87,05
	SS	10	40,71	57,65	66,10	70,89	73,76	75,49	76,63	77,28	77,66	77,91
		25	41,08	58,76	68,00	73,65	77,20	79,68	81,38	82,60	83,55	84,25
		50	40,85	58,83	68,41	74,35	78,20	80,94	82,89	84,34	85,49	86,41
		75	40,71	58,74	68,44	74,34	78,26	81,14	83,27	84,87	86,13	87,15
		100	40,49	58,54	68,19	74,19	78,22	81,21	83,41	85,02	86,37	87,40

Fonte: elaborado pelo autor (2023)

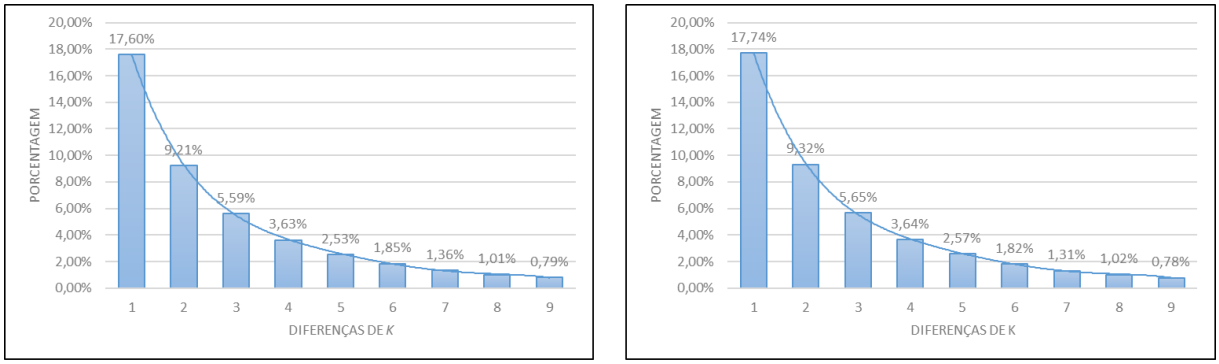
Levando-se em consideração a mediana para as estratégias de ordenação SO e SS, a Tabela 2 e a Tabela 3, respectivamente, têm os valores de 77,27% para SO e 77,47% para SS. Dessa forma, para a discussão inicial deste e dos demais modelos será considerado o parâmetro n=50 e k=5.

Isso significa que, durante a fase de teste, ao se realizarem consultas à base de conhecimento, buscam-se os 50 vetores (patentes) mais semelhantes para serem utilizados na recomendação de maneira ordenada (*ranking*) das 5 subclasses mais relevantes. Esses valores de k e n resultaram numa acurácia de 77,77% para a estratégia de ordenação SO e de 78,20% para a estratégia de ordenação SS.

Com o intuito de evidenciar a escolha da quantidade de subclasses mais relevantes (k), criou-se o Gráfico 3, que apresenta a média da diferença percentual entre as acurácias para as duas estratégias. O Gráfico 3(a) representa a estratégia SO e o Gráfico 3(b) representa a estratégia SS. Inicialmente, cada um dos valores é calculado considerando a diferença de acurácia ((k+1) – k). Com base no Gráfico 3, para a estratégia de ordenação SS e para n=10 calculam-se as diferenças de k=2 - k=1, k=3 – k=2 ... k=10 – k=9. O cálculo é repetido para os demais valores de n (25, 50, 75, 100) e, dessa forma, tem-se uma matriz com a diferença de acurácia para cada combinação de k e n.

Por fim, calcula-se a média das diferenças para cada k, ou seja, faz-se a média das diferenças ao longo das colunas, obtendo-se um valor médio por linha (k). Ao final, tem-se um vetor com a média da diferença de acurácia para cada valor de k em relação ao k anterior (k-1). Esses resultados permitem analisar o impacto médio na acurácia à medida que o número de subclasses recomendadas aumenta. Para os demais PTMs, realizou-se o mesmo procedimento.

Gráfico 3 – Média das diferenças percentuais da acurácia para o modelo *all-MiniLM-L6-v2*



Fonte: elaborado pelo autor (2023)

Analisando-se o gráfico, independentemente da estratégia utilizada, a média da diferença de acurácia tende a uma estabilização entre 5 e 6. Analisando-se as médias por estratégia, tem-se que na estratégia SO a maior diferença ocorre entre *k*=1 (média de 17,60%) e *k*=2 (média de 9,21%), uma queda de 47,67% na média das diferenças. Na estratégia SS, a maior diferença ocorre entre *k*=1 (média de 17,74%) e *k*=2 (média de 9,32%), uma queda de 47,43% na média da diferença das acurácias.

Após *k*=5, a diminuição na média das acurácias segue em ritmo menos acentuado nas duas estratégias, SO e SS. Na estratégia SO a diferença entre *k*=5 (média de 2,53%) e *k*=6 (média de 1,85%) é de 0,68%, uma queda de 27,06% na média das diferenças. Já na estratégia SS o valor da diferença entre *k*=5 (média de 2,57%) e *k*=6 (média de 1,82%) é de 0,74%, uma queda de 29,02% na média das diferenças.

Isso sugere que, a partir de *k*=5 ou *k*=6, o valor de *k* começa a se estabilizar, pois o aumento adicional de *k* não resulta em grandes mudanças nas médias das diferenças nas acurácias. Portanto, pode-se considerar que o valor de *k* tende a se estabilizar em torno de *k*=5 ou *k*=6.

5.2.1.2 Modelo all-mpnet-base-v2

O modelo *all-mpnet-base-v2* foi o que apresentou a melhor acurácia entre os PTMs utilizados, conforme os resultados apresentados na Tabela 7. Com os valores de *k*=5 e *n*=50, tem-se para a estratégia de ordenação SO uma acurácia de 78,11% e para a estratégia SS uma acurácia de 78,50%.

Tabela 7 – Acurácias para o modelo *all-mpnet-base-v2*

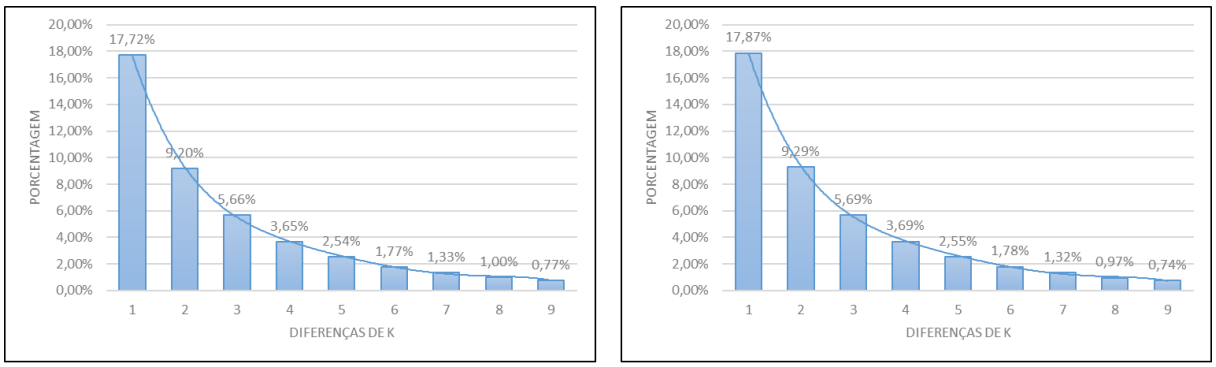
Modelo	Estratégia de ordenação	*n*	*k* =1	*k* =2	*k* =3	*k* =4	*k* =5	*k* =6	*k* =7	*k* =8	*k* =9	*k* =10
all-mpnet-base-v2	SO	10	40,80	57,56	65,82	70,60	73,54	75,39	76,52	77,24	77,67	77,94
		25	41,11	58,87	68,01	73,60	77,12	79,54	81,23	82,47	83,37	84,01
		50	40,99	58,98	68,43	74,26	78,11	80,82	82,67	84,17	85,34	86,24
		75	40,77	58,83	68,36	74,38	78,35	81,15	83,21	84,77	86,01	87,01
		100	40,63	58,68	68,30	74,36	78,35	81,27	83,39	85,02	86,29	87,34
	SS	10	40,94	57,98	66,45	71,18	73,99	75,79	76,82	77,45	77,79	77,99
		25	41,22	59,12	68,37	74,02	77,62	80,04	81,77	82,93	83,80	84,47
		50	41,06	59,20	68,70	74,58	78,50	81,25	83,17	84,71	85,86	86,73
		75	40,84	58,98	68,56	74,59	78,65	81,54	83,62	85,26	86,44	87,40
		100	40,70	58,81	68,47	74,61	78,69	81,60	83,75	85,41	86,71	87,71

Fonte: elaborado pelo autor (2023)

O Gráfico 4 apresenta as médias das diferenças percentuais das acurácias para o modelo *all-mpnet-base-v2,* onde se percebe a semelhança das curvas para as estratégias SO e SS, apresentadas nos gráficos 4(a) e 4(b), respectivamente. Analisando-se as médias por estratégia, tem-se que na estratégia SO a maior diferença ocorre entre *k*=1 (média de 17,72%) e *k*=2 (média de 9,20%), uma queda de 48,10% na média das diferenças. Já na estratégia SS a maior diferença ocorre entre *k*=1 (média de 17,87%) e *k*=2 (média de 9,29%), uma queda de 47,99% na média da diferença das acurácias.

A partir de *k*=5, a diminuição na média das acurácias segue em ritmo menos acentuado nas estratégias SO e SS. Na estratégia SO a diferença entre *k*=5 (média de 2,54%) e *k*=6 (média de 1,77%) é de 0,77%, uma queda de 30,28% na média das diferenças. Na estratégia SS a diferença entre *k*=5 (média de 2,55%) e *k*=6 (média de 1,78%) é de 0,77%, uma queda de 30,05% na média das diferenças.

Gráfico 4 – Média das diferenças percentuais da acurácia do modelo *all-mpnet-base-v2*



Fonte: elaborado pelo autor (2023)

5.2.1.3 Modelo all-distilroberta-v1

O modelo *all-distilroberta-v1* apresentou resultados semelhantes aos dois modelos anteriores, sendo pouco superior ao *all-MiniLM-L6-v2* e inferior ao *all-mpnet-base-v2*, conforme mostra a Tabela 8. Com os valores de k=5 e n=50, tem-se para a estratégia de ordenação SO uma acurácia de 77,32%, e para a estratégia SS uma acurácia de 77,67%.

Tabela 8 – Acurácias para o modelo *all-distilroberta-v1*

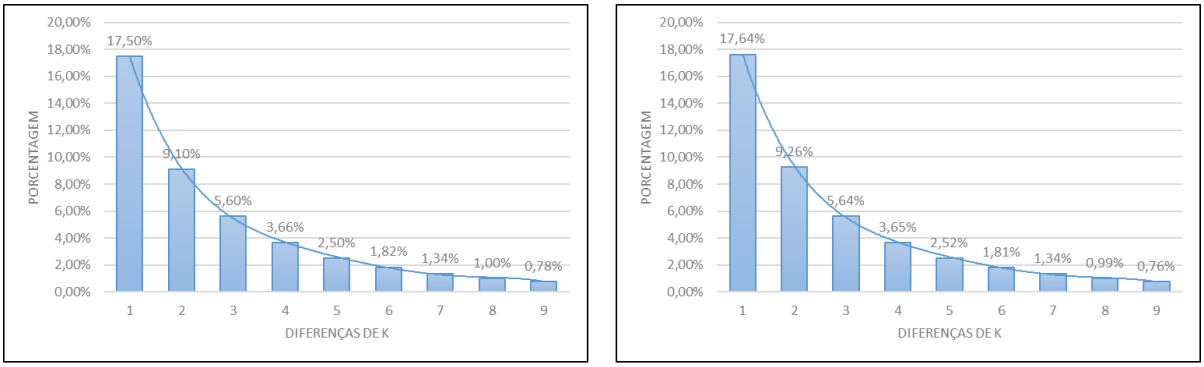
Modelo	Estratégia de ordenação	*n*	*k* =1	*k* =2	*k* =3	*k* =4	*k* =5	*k* =6	*k* =7	*k* =8	*k* =9	*k* =10
all-distilroberta-v1	SO	10	40,35	57,00	65,16	69,79	72,72	74,51	75,72	76,48	76,92	77,18
		25	40,63	58,15	67,11	72,74	76,23	78,65	80,38	81,59	82,52	83,20
		50	40,53	58,31	67,61	73,40	77,32	80,00	81,97	83,49	84,63	85,52
		75	40,35	58,19	67,66	73,56	77,54	80,27	82,35	83,97	85,21	86,19
		100	40,16	57,89	67,51	73,56	77,56	80,42	82,56	84,15	85,42	86,50
	SS	10	40,50	57,42	65,75	70,50	73,30	74,96	76,04	76,69	77,07	77,28
		25	40,70	58,36	67,53	73,16	76,75	79,21	80,88	82,10	82,98	83,62
		50	40,62	58,48	68,01	73,80	77,67	80,43	82,49	84,01	85,14	86,04
		75	40,39	58,31	67,92	73,85	77,84	80,63	82,76	84,40	85,66	86,63
		100	40,21	58,03	67,71	73,80	77,82	80,72	82,87	84,52	85,80	86,88

Fonte: elaborado pelo autor (2023)

O Gráfico 5 apresenta as médias das diferenças percentuais das acurácias para o modelo *all-distilroberta-v1* no tocante às estratégias de ordenação SO e SS, apresentadas nos gráficos 5(a) e 5(b), respectivamente. Analisando-se as médias por estratégia, tem-se que na estratégia SO a maior diferença ocorre entre k=1 (média de 17,50%) e k=2 (média de 9,10%), uma queda de 47,98% na média das diferenças. Já na estratégia SS a maior diferença ocorre entre k=1 (média de 17,64%) e k=2 (média de 9,26%), uma queda de 47,48% na média da diferença das acurácias.

A curva decresce de forma menos acentuada a partir de k=5 para as estratégias SO e SS. Na estratégia SO a diferença entre k=5 (média de 2,50%) e k=6 (média de 1,82%) é de 0,67%, uma queda de 26,97% na média das diferenças. Na estratégia SS a diferença entre k=5 (média de 2,52%) e k=6 (média de 1,81%) é de 0,70%, uma queda de 27,87% na média das diferenças.

Gráfico 5 – Média das diferenças percentuais das acurácias para o modelo *all-distilroberta-v1*



Fonte: elaborado pelo autor (2023)

5.2.1.4 Modelo en_core_web_lg

Já o modelo *en_core_web_lg* apresentou o pior resultado entre os PTMs analisados, conforme mostra a Tabela 9. Com os valores de *k*=5 e *n*=50, a estratégia de ordenação SO obteve uma acurácia de 55,95%, e a estratégia SS uma acurácia de 56,81%. Conforme declarado anteriormente, esse PTM é caracterizado como não contextual, demonstrando ser inadequado pelo menos para a tarefa de recomendação ordenada de subclasses.

Tabela 9 – Acurácias para o modelo *en_core_web_lg*

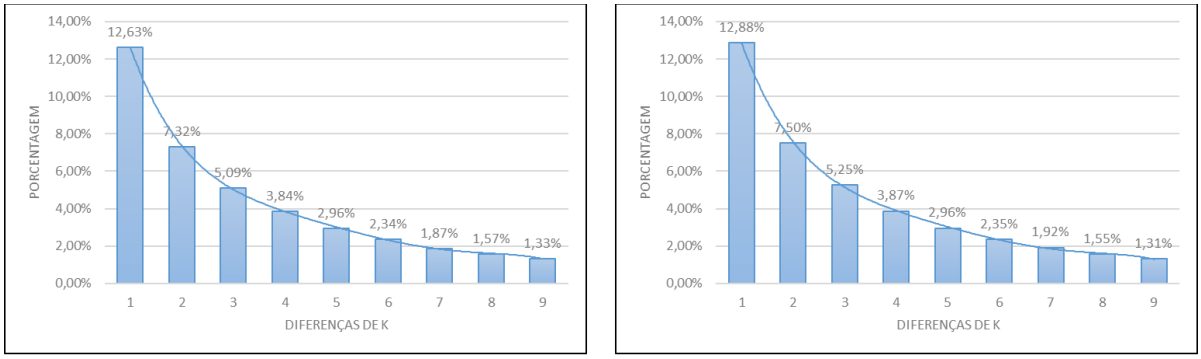
Modelo	Estratégia de ordenação	*n*	*k* =1	*k* =2	*k* =3	*k* =4	*k* =5	*k* =6	*k* =7	*k* =8	*k* =9	*k* =10
en_core_web_lg	SO	10	25,50	36,83	43,26	47,49	50,86	53,43	55,51	57,16	58,49	59,38
		25	26,46	38,70	45,85	50,95	54,73	57,55	59,69	61,39	62,69	63,85
		50	26,34	39,27	46,87	52,06	55,96	59,05	61,48	63,43	65,18	66,68
		75	26,13	39,36	47,08	52,55	56,57	59,69	62,15	64,19	65,90	67,40
		100	25,90	39,31	47,00	52,46	56,62	59,80	62,40	64,44	66,22	67,81
	SS	10	26,31	38,08	44,91	49,46	52,65	55,01	56,78	58,13	59,12	59,80
		25	26,86	39,37	46,79	51,97	55,84	58,77	61,09	62,97	64,46	65,68
		50	26,61	39,82	47,45	52,83	56,82	59,93	62,46	64,55	66,29	67,77
		75	26,36	39,73	47,54	53,09	57,21	60,39	62,94	65,04	66,82	68,39
		100	26,07	39,60	47,41	53,01	57,20	60,45	63,00	65,18	66,95	68,56

Fonte: elaborado pelo autor (2023)

O Gráfico 6 apresenta as médias das diferenças percentuais das acurácias para o modelo *en_core_web_lg* para as estratégias de ordenação SO e SS, apresentadas nos gráfico 6(a) e 6(b), respectivamente. O modelo obteve a maior diferença na média para a estratégia entre *k*=1 (média de 12,63%) e *k*=2 (média de 7,32%), uma queda de 42,07% na média das diferenças. Já para a estratégia SS, a maior diferença ocorre entre *k*=1 (média de 12,88%) e *k*=2 (média de 7,50%), uma queda de 41,75% na média da diferença das acurácias.

Assim como nos demais PTMs, a curva decresce de forma menos acentuada a partir de *k*=5 para as estratégias SO e SS. Na estratégia SO a diferença entre *k*=5 (média de 2,96%) e *k*=6 (média de 2,35%) é de 0,61%, uma queda de 20,69% na média das diferenças. Na estratégia SS a diferença entre *k*=5 (média de 2,96%) e *k*=6 (média de 2,35%) é de 0,62%, uma queda de 20,90% na média das diferenças.

Gráfico 6 – Média das diferenças percentuais das acurácias para o modelo *en_core_web_lg*



Fonte: elaborado pelo autor (2023)

5.2.2 Comparação do modelo atual com o modelo de redes neurais

O propósito desta seção é comparar os resultados obtidos pelo modelo proposto utilizando PTMs e as estratégias específicas de *ranking* com diferentes redes neurais clássicas de DL, com o intuito de analisar o desempenho do modelo.

Vale ressaltar que as redes neurais clássicas utilizadas nessa avaliação fizeram parte da primeira versão do modelo proposto (Apêndice C) e que, após diferentes testes, algumas limitações foram percebidas. Entre essas limitações estão a alta demanda por recurso computacional para lidar com grandes volumes de dados bem como a necessidade frequente de atualização dos modelos treinados (*fine-tuning*). Já a versão final do modelo possui um custo linear, visto que à medida que novas patentes são avaliadas e classificadas pelo examinador, estas possuem seus *embeddings* gerados, sendo indexadas na base de conhecimento. A partir disso, passam a integrar a próxima recomendação ordenada de subclasses. Ademais, quando determinada patente é incorporada à base de conhecimento, ocorre também a atualização do KG, o que confere ao modelo dinamicidade e capacidade de evolução temporal.

Para a comparação com o modelo proposto, foram utilizadas as arquiteturas de redes neurais MLP, CNN e LSTM. No contexto do presente trabalho, essa comparação tem como objetivo avaliar a eficácia do modelo proposto utilizando-se PTMs e estratégias de *ranking* de

subclasses de patentes em comparação com arquiteturas tradicionais de redes neurais na tarefa de classificação com capacidade de produção de *rankings*. Ademais, essa comparação visa determinar a melhor abordagem para a tarefa principal desta tese, considerando a disponibilidade de dados e os recursos computacionais.

No Quadro 23 estão dispostas as configurações utilizadas nos testes com as redes neurais e os PTMs. O conjunto de dados foi reduzido, sendo composto por patentes dos anos de 2012, 2013 e 2014 e, ao final, dividido em dois outros conjuntos, treinamento e teste. Armazenaram-se as patentes em uma tabela contendo o título, o resumo, a lista de subclasses e um tipo, este para representar se a patente deveria ser utilizada na etapa de treinamento ou de teste.

Para o cenário de estudo, foram identificadas as 50 subclasses de patentes mais frequentes extraídas do conjunto de dados total (composto por quase 2 milhões de patentes). Realizou-se a análise de todas as patentes, e as subclasses associadas foram contabilizadas em uma estrutura que contém o código da subclasse e a frequência, ou seja, a quantidade de patentes que pertencem à subclasse, sendo então selecionadas as 50 mais frequentes.

Como resultado final desse processo, o conjunto de treinamento foi composto por 40 mil patentes, e o conjunto de teste por 10 mil patentes, ou seja, 1.000 (mil) patentes de cada uma das 50 subclasses. Dessas 1.000 patentes, 800 (oitocentas) serviram para o treinamento e 200 (duzentas) para a etapa de teste. Tanto o conjunto de treinamento quanto o conjunto de teste são compostos por duas colunas, uma indicando a subclasse (atributo meta) e outra indicando o texto da patente, ou seja, a concatenação de título e resumo (Apêndice D).

Quadro 23 – Configuração do conjunto de dados para comparação entre modelos

Modelo	Remoção de *Stopwords*	Subclasses	Dados	Número de épocas	Conjunto de treinamento	Conjunto de teste	Estratégia de ordenação
all-MiniLM-L6-v2	Sim	50	Título/Resumo	-	40.000	10.000	SO e SS
all-mpnet-base-v2	Sim	50	Título/Resumo	-	40.000	10.000	SO e SS
all-distilroberta-v1	Sim	50	Título/Resumo	-	40.000	10.000	SO e SS
en_core_web_lg	Sim	50	Título/Resumo	-	40.000	10.000	SO e SS
MLP	Sim	50	Título/Resumo	100	40.000	10.000	-
CNN	Sim	50	Título/Resumo	100	40.000	10.000	-
LSTM	Sim	50	Título/Resumo	100	40.000	10.000	-

Fonte: elaborado pelo autor (2023)

No que se refere ao número de épocas, este somente é aplicado às redes neurais na fase de treinamento. Uma época representa a passagem por todo o conjunto de dados. Ademais, vale mencionar que para o treinamento ocorre também a validação do que foi aprendido em

determinada época, ou seja, do total de instâncias do treinamento, um percentual é considerado (utilizou-se o total de 10%).

Já os PTMs foram avaliados com base nas duas estratégias de ordenação propostas na tese. Todavia, para efeitos de comparação com as redes neurais, utilizou-se a estratégia da soma dos *scores* (SS), a qual não se aplica para os modelos de redes neurais CNN, LSTM e MLP.

O Quadro 24 mostra os melhores resultados na avaliação da recomendação ordenada (*ranking*) de subclasses de patentes, com *n*=50 gerados pelos PTMs, *all_Mini-LM-L6-v2*, *all-mpnet-base-v2*, *all-diltilroberta-v1* e *en_core_web_lg*, e por meio das redes neurais CNN, LSTM e MLP.

Quadro 24 – Comparação das abordagens utilizadas

PTMs/ANNs	*Stopwords*	Lematização	Dados	Número de épocas	Estratégia de ordenação	Acurácia *k* =1	Acurácia *k* =3	Acurácia *k* =5
all-mpnet-base-v2	Sim	Não	Título/Resumo	-	SS	57,02	83,01	90,87
all-MiniLM-L6-v2	Sim	Não	Título/Resumo	-	SS	56,82	82,52	90,77
all-distilroberta-v1	Sim	Não	Título/Resumo	-	SS	55,83	81,68	90,32
en_core_web_lg	Sim	Não	Título/Resumo	-	SS	42,29	69,22	80,68
CNN	Sim	Sim	Título/Resumo	100	-	39,87	64,97	77,08
LSTM	Sim	Sim	Título/Resumo	100	-	38,19	62,48	74,43
MLP	Sim	Sim	Título/Resumo	100	-	39,42	62,50	73,58

Fonte: elaborado pelo autor (2023)

Para elaboração do Quadro 24, observaram-se as configurações das redes neurais que obtiveram melhor resultado e, a partir disso, foram selecionados os PTMs no desenvolvimento final do modelo proposto considerando os mesmos parâmetros. A coluna “Estratégia de ordenação” não se aplica às arquiteturas de redes neurais, pois a saída destas é formada por um vetor de probabilidades com dimensionalidade igual ao número de subclasses, em que o *ranking* é constituído ordenando-se do maior para o menor valor, ou seja, das subclasses que possuem maior relevância para as que possuem menor relevância. Sendo assim, para a recomendação basta indicar os valores até determinado *k*. Já a coluna “Número de épocas” refere-se somente às redes neurais tradicionais de DL.

Os resultados indicam a análise da acurácia do *ranking* para as *k* subclasses mais relevantes com os valores 1, 3 e 5 e *n*=50 (o *n* aplica-se somente aos PTMs). Sendo assim, considerando-se uma patente de entrada para a etapa de teste apresentada ao modelo e também as redes neurais tradicionais, levando-se em conta determinado valor de *k* (valores utilizados 1, 3 e 5), tem-se a acurácia para cada uma das avaliações do Quadro 24.

Os melhores resultados foram obtidos para os modelos de similaridade vetorial *all-mpnet-base-v2*, *all-MiniLM-L6-v2*, *all-distilroberta-v1*, em conjunto com a estratégia de

ordenação de soma dos *scores* (SS). O modelo *all-mpnet-base-v2* teve um percentual de acertos de 90,87% para *k=5*.

Na Tabela 10, para efeitos de comparação entre as redes neurais e o PTM de melhor resultado *all-mpnet-base-v2* (mpnet), tem-se a acurácia do *ranking* para as dez subclasses (*top* 10) mais relevantes. Nesse sentido, considerando-se determinada patente na etapa de teste apresentada ao modelo de classificação e levando-se em conta a primeira subclasse sugerida (*ranking* igual a 1), tem-se uma acurácia para cada uma das abordagens. O mesmo ocorre para as demais posições do *ranking*.

Como os resultados obtidos para as três redes neurais são próximos, para efeitos de análise serão discutidos os resultados da rede CNN, que obteve um desempenho levemente superior às demais com o método *all-mpnet-base-v2*.

Tabela 10 – Comparação das acurácias entre as redes neurais e o modelo *all-mpnet-base-v2*

	Acurácia			
Ranking	**CNN**	**LSTM**	**MLP**	**mpnet**
1	39,87%	38,19%	39,42%	57,02%
2	55,55%	53,00%	53,48%	74,14%
3	64,97%	62,48%	62,50%	83,01%
4	71,90%	69,40%	68,58%	87,99%
5	77,08%	74,43%	73,58%	90,87%
6	80,64%	78,24%	76,90%	92,77%
7	83,28%	81,00%	79,97%	94,16%
8	85,55%	83,22%	82,10%	94,98%
9	87,35%	85,20%	84,26%	95,51%
10	88,80%	87,03%	85,91%	95,86%

Fonte: elaborada pelo autor (2023)

Nesse sentido, percebe-se pela Tabela 10 que a acurácia obtida considerando a primeira a recomendação de 1 (uma) subclasse é de 39,87% para a CNN e o 57,02% para o modelo *mpnet*. Pensando-se em um cenário mais próximo da aplicação real do modelo, pelo menos 5 (cinco) subclasses seriam ofertadas para a análise de um examinador. Sendo assim, torna-se mais provável que entre as 5 (cinco) primeiras subclasses exista pelo menos uma subclasse que poderia ser vinculada à patente analisada. Essa situação em particular atingiu uma acurácia de 77,08% para a rede CNN e 90,87% para o método *mpnet*.

Calculando-se a variação percentual para a primeira posição do *ranking,* tem-se um valor de 43,01% a favor do *mpnet* em relação à CNN; já para a quinta posição do *ranking*, o mpnet obteve um aumento de 17,89%. Percebe-se que à medida que o *k* cresce, a diferença

entre a variação diminui. Apesar de se esperar isso, a oferta de muitas subclasses pode dificultar o processo e impactar na tomada de decisão de determinado examinador.

As próximas seções apresentam uma análise individual de algumas patentes com o objetivo de clarificar o *ranking* utilizando os PTMs e as redes neurais tradicionais. Conforme demonstrado na seção 4.3.4, para cada patente apresentada ao modelo obtém-se uma lista de subclasses ordenadas pela relevância dessas subclasses. De modo geral, tal relevância pode ser entendida como uma medida da importância da subclasse, permitindo o *ranking*.

5.2.2.1 Cenário para a patente nº US08472379

Para clarificar o funcionamento do processo de ordenação das subclasses (*ranking*), o Quadro 25 apresenta um exemplo de patente utilizada nos testes com as redes neurais CNN, LSTM e MLP. Essa patente também será utilizada como exemplo para os PTMs que fazem parte da avaliação principal do modelo proposto nesta tese. A patente nº US08472379 do ano de 2013 que consta no conjunto de teste possui uma classe H04 e duas subclasses, representadas por H04J e H04W, sem qualquer tipo de ordenação. A primeira página do documento da patente original encontra-se no Anexo A.

Quadro 25 – Patente de exemplo

Campos	**Conteúdo**
Título	Mobile station radio base station communication control method and mobile communication system
Resumo	A mobile station according to the present invention includes a packet discarder unit configured to discard a packet in an uplink transmission buffer after assigning a sequence number to the packet when a predetermined condition is met.
Subclasse	H04J, H04W (ordem que aparece no conjunto de dados sem qualquer indicação de relevância)

Fonte: elaborado pelo autor (2023)

O resultado da execução da etapa de teste da patente de exemplo é apresentado no Quadro 26. As redes neurais CNN, LSTM e MLP apresentam a ordenação para as 10 (dez primeiras) subclasses da patente de exemplo. A rede CNN recomendou a subclasse H04W na primeira posição do *ranking*, com uma probabilidade de 40,08%. Já a probabilidade de acerto da subclasse H04J foi de 20,47% na segunda posição do *ranking*. A probabilidade é gerada para as 50 posições de cada patente. Dessa forma, a soma das 50 probabilidades da patente no *ranking* deve ser igual a 1 (100%).

A rede LSTM obteve uma probabilidade de acerto de 41,88% para a subclasse H04W (em azul), ficando na primeira posição do *ranking*, e 17,16% para a subclasse H04J (em verde),

terceira posição do *ranking*. Já a rede MLP obteve uma probabilidade de acerto de 77,60% para a subclasse H04W, na primeira posição do *ranking*, e a subclasse H04J ficou na segunda posição, com 17,91% de probabilidade de acerto.

Quadro 26 – *Ranking* com as 10 subclasses mais relevantes para as redes neurais

Ranking	**CNN**		**LSTM**		**MLP**	
	Subclasse	Probabilidade	Subclasse	Probabilidade	Subclasse	Probabilidade
1	H04W	40,08	H04W	41,88	H04W	77,60
2	H04J	20,48	H04B	23,02	H04J	17,91
3	H04B	16,16	H04J	17,16	H04B	1,49
4	H04L	12,14	H04L	8,90	H04L	0,87
5	H04M	6,89	H04M	7,95	C12P	0,63
6	G01R	1,66	G01R	0,80	G01R	0,49
7	G06F	1,09	G08B	0,11	A63F	0,25
8	G08B	0,46	G06F	0,10	C12Q	0,18
9	E21B	0,17	H03K	0,05	H01L	0,15
10	H03K	0,12	H04N	0,01	H04M	0,12

Fonte: elaborado pelo autor (2023)

Com base nos PTMs, tem-se a ordenação para as *k* primeiras subclasses da patente de exemplo. O resultado levou em consideração *k*=10 e *n*=50. Para essa avaliação e as que constam nas próximas duas seções, como existem menos subclasses para o conjunto de dados que forma a base de conhecimento (50 ao todo), o *k* indicará um valor máximo de subclasses a serem apresentadas para cada um dos PTMs. A frequência, isto é, o número de documentos retornados que mencionam a subclasse, será de no máximo 50, visto que em determinado caso o número de subclasses recomendadas pode ser superior a 10 (o Quadro 28 é um exemplo).

Para calcular a relevância da subclasse (coluna Relevância (%)), deve-se considerar a frequência com que a subclasse é mencionada nos documentos (patentes) retornados, sendo essa frequência dividida por 50 (número total de documentos retornados). Menciona-se que nessa avaliação e nas demais utilizando PTMs tem-se o conceito de relevância em vez de probabilidade. Isso ocorre porque nos 50 documentos retornados nem todas as subclasses estarão presentes. Por outro lado, nas redes neurais a saída é sempre um vetor com as 50 subclasses (considerando este cenário de estudo), em que cada posição indica a probabilidade da respectiva subclasse. Analisando-se o Quadro 27 para o PTM *all-mpnet-base-v2*, a subclasse H04W (em azul) possui uma relevância de 44,00%. Essa subclasse e a subclasse H04J, com uma relevância de 26,00%, fazem parte do conjunto de subclasses da patente de entrada (Quadro 25).

Quadro 27 – *Ranking* com as *k* subclasses mais relevantes para os PTMs *all-MiniLM-L6-v2* e *all-mpnet-base-v2*

Ranking	***all-MiniLM-L6-v2***			***all-mpnet-base-v2***		
	Subclasse	Frequência	%	Subclasse	Frequência	%
1	H04W	22	44,00	H04W	21	42,00
2	H04J	13	26,00	H04J	9	18,00
3	H04B	8	16,00	H04B	8	16,00
4	H04L	4	8,00	H04L	7	14,00
5	H04M	1	2,00	G01R	2	4,00
6	H03M	1	2,00	H04M	1	2,00
7	G01R	1	2,00	H03M	1	2,00
8				G06F	1	2,00

Fonte: elaborado pelo autor (2023)

Considerando os resultados apresentados no Quadro 27 e no Quadro 28, percebe-se que a subclasse H04W (em azul) aparece na primeira posição em todos os PTMs, enquanto a subclasse H04J (em verde) aparece sempre na segunda posição. Para os modelos *all-MiniLM-L6-v2*, *all-mpnet-base-v2* e *all-distilroberta-v1*, é possível verificar que os resultados são mais adequados quando comparados com o modelo *en_core_web_lg*, visto que a maior parte das subclasses sugeridas pertence à classe H04. Pode-se interpretar que o conjunto de documentos recuperados possui uma melhor representação semântica, levando-se em conta a patente de entrada. Já o modelo *en_core_web_lg* recuperou documentos que produziram uma recomendação com mais de 10 classes. Considerando a soma da frequência, as subclasses foram mencionadas em 46 documentos, o que totaliza uma relevância acumulada de 92%.

Quadro 28 – *Ranking* com as *k* subclasses mais relevantes para os PTMs *all-distilroberta-v1* e *en_core_web_lg*

Ranking	***all-distilroberta-v1***			***en_core_web_lg***		
	Subclasse	Frequência	%	Subclasse	Frequência	%
1	H04W	22	44,00	H04W	16	32,00
2	H04J	13	26,00	H04J	8	16,00
3	H04B	11	22,00	H04M	6	12,00
4	G01R	2	4,00	H04B	5	10,00
5	H04M	1	2,00	G08B	4	8,00
6	H04L	1	2,00	H04L	3	6,00
7				H04N	1	2,00
8				H01R	1	2,00
9				G11C	1	2,00
10				G09G	1	2,00

Fonte: elaborado pelo autor (2023)

Em relação à subclasse que aparece na primeira posição H04W (em azul), deve-se considerá-la como sugestiva à invenção na sua totalidade ou como o principal conceito inventivo utilizado, levando-se em conta o título e o resumo da patente. A experiência e o conhecimento do estado da arte por parte dos examinadores que realizam a classificação de documentos de patentes podem exercer influência na forma como esses documentos são categorizados, permitindo que um documento de patente seja classificado em uma ou mais subclasses.

Analisando-se os resultados dos PTMs e das redes neurais, nesse caso específico, verifica-se que todos são similares. Sendo assim, é pertinente investigar se o esforço para determinar a arquitetura de rede neural mais adequada é um fator determinante ou irrelevante para o modelo proposto. Em outras palavras, é preciso avaliar se, independentemente da arquitetura de rede neural, o modelo proposto é capaz de fornecer resultados que auxiliem na tomada de decisão dos examinadores. Todavia, os resultados para esse cenário de estudo sugerem que arquiteturas do tipo *transformers* e a utilização de estratégias específicas de *ranking* promovem, no geral, resultados mais adequados.

5.2.2.2 Cenário para a patente nº US08394786

A patente US08394786 foi utilizada para analisar o cenário em que na extração dos dados não foi observada a completude da parte textual. Essa patente foi publicada no ano de 2013 e tem como subclasses C07D e A61K, como mostra o Quadro 29. Utiliza-se esse exemplo

para demonstrar como o modelo se comporta com a subtração de parte do *abstract* em que se tenha menos conteúdo textual para se chegar a resultados consistentes. A primeira página do documento de patente original encontra-se no Anexo B.

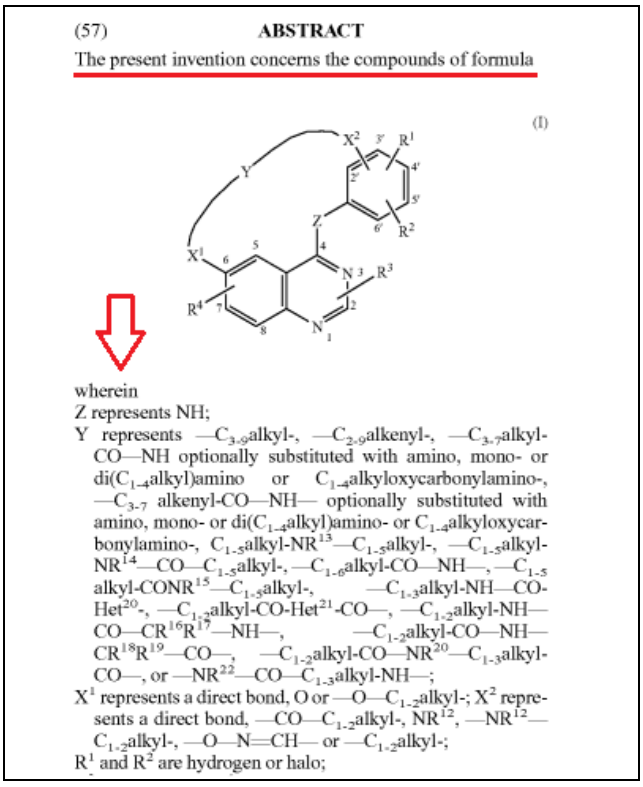
Quadro 29 – Patente US08394786

Campos	**Conteúdo**
Título	quinazoline derivatives
Resumo	the present invention concerns the compounds of formula…
Subclasse	C07D, A61K (ordem que aparece no conjunto de dados sem qualquer indicação de relevância)

Fonte: elaborado pelo autor (2023)

A Figura 35 mostra a parte não inserida no conjunto de dados da USPTO-2M, visto que a extração do texto do *abstract* da patente US08394786 foi interrompida ao encontrar uma fórmula; o restante do texto continua após isso, dessa forma o *abstract* da patente encontra-se incompleto. A parte sublinhada na Figura 35 mostra a parte extraída, e a seta indica a continuação do texto do *abstract* que não foi incorporado ao conjunto de dados USPTO-2M.

Figura 35 – *Abstract* da patente US08394786



(57) **ABSTRACT**

The present invention concerns the compounds of formula

(I)

wherein

Z represents NH;

Y represents —C_{3-9}alkyl-, —C_{2-9}alkenyl-, —C_{3-7}alkyl-CO—NH optionally substituted with amino, mono- or di(C_{1-4}alkyl)amino or C_{1-4}alkyloxycarbonylamino-, —C_{3-7} alkenyl-CO—NH— optionally substituted with amino, mono- or di(C_{1-4}alkyl)amino- or C_{1-4}alkyloxycarbonylamino-, C_{1-5}alkyl-NR^{13}—C_{1-5}alkyl-, —C_{1-5}alkyl-NR^{14}—CO—C_{1-5}alkyl-, —C_{1-6}alkyl-CO—NH—, —C_{1-5}alkyl-$CONR^{15}$—C_{1-5}alkyl-, —C_{1-3}alkyl-NH—CO-Het^{20}-, —C_{1-2}alkyl-CO-Het^{21}-CO—, —C_{1-2}alkyl-NH—CO—$CR^{16}R^{17}$—NH—, —C_{1-2}alkyl-CO—NH—$CR^{18}R^{19}$—CO—, —C_{1-2}alkyl-CO—NR^{20}—C_{1-3}alkyl-CO—, or —NR^{22}—CO—C_{1-3}alkyl-NH—;

X^1 represents a direct bond, O or —O—C_{1-2}alkyl-; X^2 represents a direct bond, —CO—C_{1-2}alkyl-, NR^{12}, —NR^{12}—C_{1-2}alkyl-, —O—N=CH— or —C_{1-2}alkyl-;

R^1 and R^2 are hydrogen or halo;

Fonte: elaborado pelo autor (2023)

Considerando-se o resultado apresentado no Quadro 30, percebe-se que a subclasse C07D (em azul) aparece na primeira posição nas redes neurais CNN e LSTM, ao passo que a subclasse A61K (em verde) aparece na segunda posição para a CNN e na terceira posição para a LSTM. A rede MLP indicou a subclasse A61K (em verde) na primeira posição e a subclasse C07D (em azul) na segunda posição.

A rede CNN classificou como resposta correta a subclasse C07D, com um percentual de acertos de 46,94%. Já a probabilidade de acerto da subclasse A61K foi de 21,64%. Para a rede LSTM, as subclasses C07D e A61K tiveram uma probabilidade de 58,71% e de 12,40%, respectivamente. Já a rede MLP obteve uma probabilidade de acerto de 73,50% para a subclasse C07D, e a subclasse A61K teve 10,05% de probabilidade de acerto.

Quadro 30 – Resultados dos testes com as redes neurais para a patente US08394786

Ranking	**CNN**		**LSTM**		**MLP**	
	Subclasse	Probabilidade	Subclasse	Probabilidade	Subclasse	Probabilidade
1	C07D	46,94	C07D	58,71	A61K	73,50
2	A61K	21,64	C07C	15,30	C07D	10,05
3	A01N	19,13	A61K	12,40	A01N	9,28
4	C07C	8,74	A01N	11,32	C07H	3,50
5	C07H	1,15	G01N	1,38	A61N	0,92
6	C07K	0,68	C07H	0,27	C12Q	0,52
7	C12P	0,31	B32B	0,26	C07K	0,35
8	B01D	0,28	C07K	0,12	A61M	0,34
9	G01N	0,15	B01D	0,08	G01N	0,30
10	C12Q	0,14	C12Q	0,05	A61F	0,22

Fonte: elaborado pelo autor (2023)

Com base nos PTMs, tem-se a ordenação para as *k* subclasses considerando-se a patente US08394786 como entrada. O resultado levou em conta *k=10* e *n=50*. Todavia, o número de subclasses varia de acordo com o PTM utilizado, como se pode observar nos quadros 31 e 32.

Quadro 31 – *Ranking* com as *k* subclasses mais relevantes da patente US08394786 para os PTMs *all-miniLM-L6-v2* e *all-mpnet-base-v2*

Ranking	***all-MiniLM-L6-v2***			***all-mpnet-base-v2***		
	Subclasse	Frequência	Relevância (%)	Subclasse	Frequência	Relevância (%)
1	C07D	22	44,00	C07D	23	46,00
2	A61K	13	26,00	A61K	14	28,00
3	A01N	12	24,00	A01N	11	22,00
4	C07C	3	6,00	C07C	2	4,00

Fonte: elaborado pelo autor (2023)

Considerando-se os resultados apresentados no Quadro 31 e no Quadro 32, verifica-se que a subclasse C07D (em azul) aparece na primeira e a subclasse A61K (em verde) aparece na segunda posição em todos os MPTs. Com exceção do MPT *en_core_web_lg*, em que a primeira e a segunda posições possuem relevâncias bem próximas, nos demais modelos verifica-se uma maior importância para a subclasse C07D.

Outro ponto se refere à soma das acurácias para as duas primeiras subclasses (primeira e segunda posições do *ranking*), sendo de 70,00% para o método *all-MiniLM-L6-v2*, 74,00% no método *all-mpnet-base-v2*, 78,00% no método *all-distilroberta-v1* e de 32,00% para o método *en_core_web_lg*. Percebe-se que nos métodos contextualizados as duas primeiras subclasses possuem acurácias acima dos 69%, enquanto no método *en_core_web_lg* a soma das duas primeiras relevâncias fica em 32,00%. Isso denota uma menor relevância dos documentos recuperados, resultando em um número maior de subclasses que não estão associadas à patente de entrada. Esse comportamento pode impactar na escolha da classificação adequada por parte de um examinador, dificultando o trabalho dele.

Os resultados obtidos para a patente US08394786 tanto para as redes neurais quanto para os PTMs são satisfatórios. Ressalta-se que em todos os testes realizados, nesse cenário em particular, a classificação foi feita adequadamente, sugerindo através do *ranking* as subclasses nas primeiras posições mesmo com a ausência da maior parte do texto do resumo da patente.

Quadro 32 – *Ranking* com as *k* subclasses mais relevantes da patente US08394786 para os PTMs *all-distilroberta-v1* e *en_core_web_lg*

Ranking	***all-distilroberta-v1***			***en_core_web_lg***		
	Subclasse	Frequência	Relevância (%)	Subclasse	Frequência	Relevância (%)
1	C07D	26	52,00	C07D	9	18,00
2	A61K	13	26,00	A61K	7	14,00
3	A01N	6	12,00	C07C	6	12,00
4	C07C	4	8,00	C07H	5	10,00
5	C12Q	1	2,00	A01N	5	10,00
6				G01N	4	8,00
7				C12P	3	6,00
8				C07K	3	6,00
9				C12N	2	4,00
10				H03M	1	2,00

Fonte: elaborado pelo autor (2023)

5.2.2.3 Cenário para a patente nº USPP022862

A patente USPP022862 foi selecionada como cenário de estudo por não possuir o título da patente no conjunto de dados da USPTO-2M. O Quadro 33 apresenta os dados da patente extraída do conjunto de dados, tratando-se de uma solicitação de um cultivar, mais especificamente da planta gérbera ("*garoran*") para o ano de publicação de 2012. Considerando a patente disponível na *web*, o título ausente no conjunto de dados USPTO-2M é "GERBERA PLANT NAMED GARORAN". A primeira página da patente consta no Anexo C.

Esse cenário se mostra interessante para verificar o comportamento do modelo proposto quando da ausência de um dos elementos textuais da patente, nesse caso a ausência do título que representa conteúdo relevante na descrição de uma patente.

Quadro 33 – Patente USPP022862

Campos	**Conteúdo**
Título	…
Resumo	a new and distinct cultivar of plant named garoran characterized by its compact upright and uniformly mounding plant habit freely flowering habit orange and yellow bi colored ray florets upright and strong scapes and good garden performance
Subclasse	A01H

Fonte: elaborado pelo autor (2023)

No Quadro 34, tem-se os resultados dos testes com as redes neurais. A patente USPP022862 foi utilizada como entrada para as redes neurais, e a subclasse atribuída para essa

patente é a A01H. Nos resultados, essa subclasse não se encontra entre as dez subclasses mais relevantes do *ranking*. Para a rede neural CNN, a subclasse A01H se encontra na 26ª (vigésima sexta posição), para a rede LSTM está na 12ª (décima segunda posição) e para a rede MLP está na 18ª (décima oitava posição).

Quadro 34 – Resultados dos testes com redes neurais para a patente USPP022862

Ranking	**CNN**		**LSTM**		**MLP**	
	Subclasse	Probabilidade	Subclasse	Probabilidade	Subclasse	Probabilidade
1	C07D	42,59	C07D	88,59	A01N	69,53
2	A61K	21,16	C07C	4,61	C07D	15,37
3	A01N	19,30	A01N	3,83	C07C	8,08
4	C07C	12,51	A61K	2,91	C07H	2,70
5	C07H	1,46	G01N	0,05	A61K	1,77
6	B01D	0,67	C07H	0,00	C12P	0,84
7	C07K	0,58	B32B	0,00	B01D	0,36
8	C12P	0,47	C07K	0,00	B32B	0,23
9	G01N	0,26	B01D	0,00	C07K	0,21
10	C12Q	0,22	C12P	0,00	C12N	0,21

Fonte: elaborado pelo autor (2023)

Assim como nos dois cenários anteriores, os testes com os PTMs para a patente USPP022862 consideraram o *k=10* e o *n=50*. Levando em conta o *ranking*, o número de subclasses retornadas foi o mesmo para os diferentes PTMs. A frequência foi de 50, com relevância de 100% para a subclasse A01H em todos os métodos utilizados, como apresentado no Quadro 35.

Quadro 35 – Resultado para a patente USPP022862 com os PTMs

Ranking	***all-MiniLM-L6-v2***			***all-mpnet-base-v2***		
	Subclasse	Frequência	Relevância (%)	Subclasse	Frequência	Relevância (%)
1	A01H	50	100,00	A01H	50	100,00

Ranking	***all-distilroberta-v1***			***en_core_web_lg***		
	Subclasse	Frequência	Relevância (%)	Subclasse	Frequência	Relevância (%)
1	A01H	50	100,00	A01H	1	100,00

Fonte: elaborado pelo autor (2023)

Basicamente, isso ocorreu devido à similaridade dos 50 documentos retornados, todos mencionando patentes relacionadas a “cultivares”. Em todos os documentos retornados, a subclasse A01H estava associada. Tal uniformidade de resultados indica que os MPTs foram bem-sucedidos ao identificar e classificar corretamente essa subclasse em particular, levando

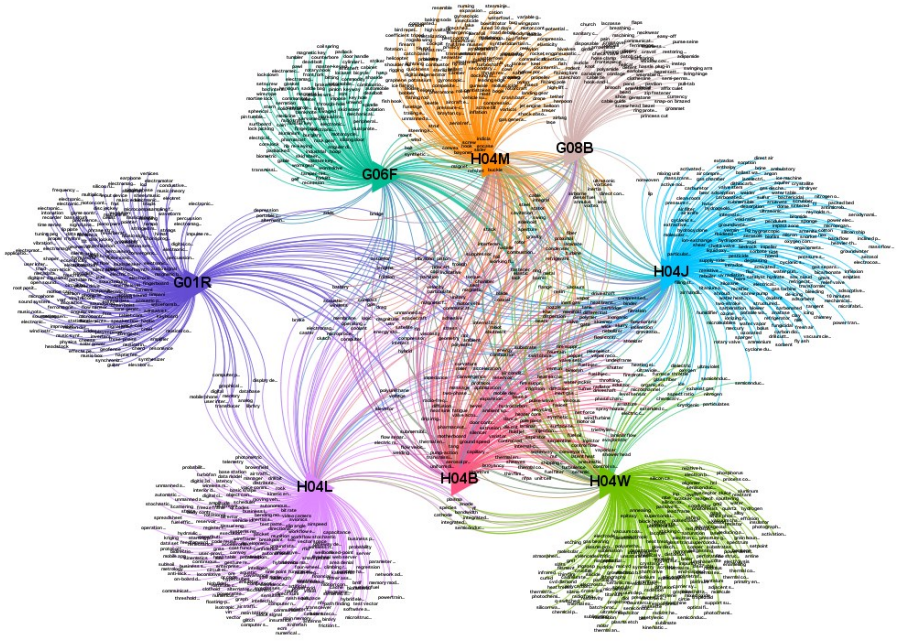
em conta somente o resumo fornecido como entrada. Por outro lado, os resultados gerados pelas redes neurais indicam uma dificuldade em mapear adequadamente o contexto de determinado domínio, produzindo *embeddings* que, nesse cenário, não foram capazes de analisar adequadamente o texto da patente de entrada.

5.3 GRAFO DE CONHECIMENTO

O presente trabalho utiliza técnicas de NLP na extração de tópicos (nós) e relacionamentos (arestas) entre estes e as subclasses, visando a geração de um grafo de conhecimento que representa o domínio de patentes no cenário de estudo. A ideia geral consiste na extração de tópicos dos documentos de patentes, associando-os às subclasses da patente em questão. Objetiva, ainda, a geração do KG, de modo que este sirva de elemento importante na explicitação do *ranking* de subclasse de patentes, de modo a facilitar a explicação e a visualização dos resultados.

Na Figura 36, são apresentadas 8 (oito) subclasses de patentes, indicadas pelos códigos H04W, H04J, H04B, H04L, H04M, G01R, G06F e G08B, com os seus respectivos conceitos.

Figura 36 – Grafo de conhecimento

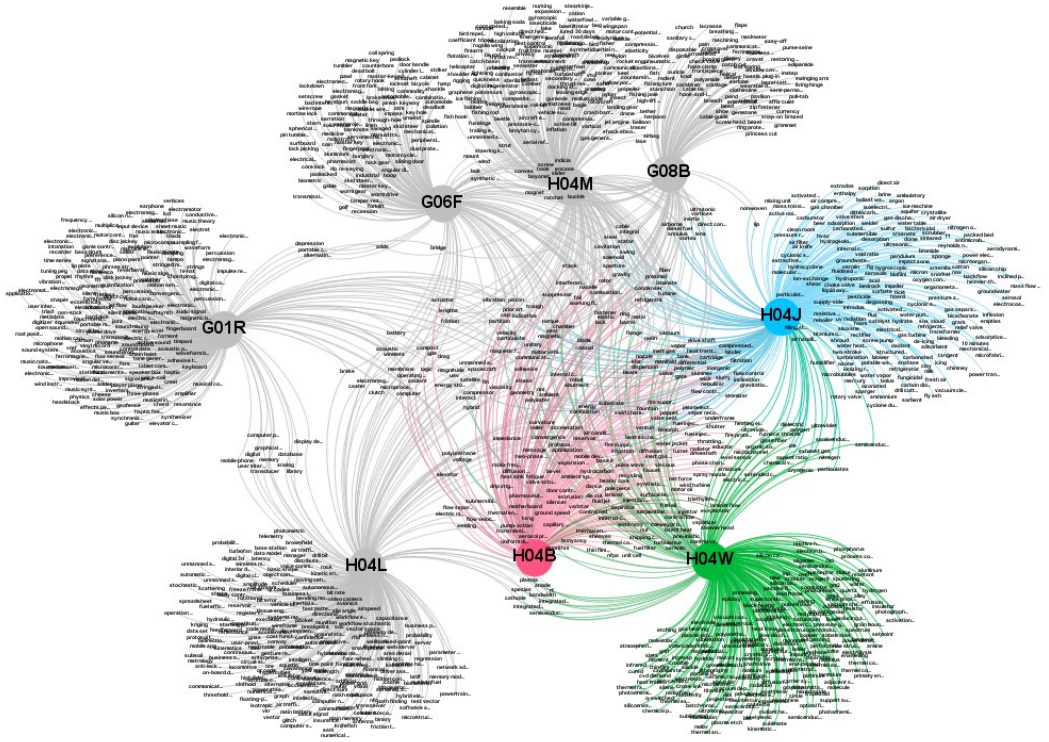


Fonte: elaborado pelo autor (2023)

O KG permite analisar as relações entre as subclasses de patentes conectadas aos seus tópicos com os respectivos pesos. O peso de cada aresta é determinado pela frequência com que um dado tópico se conecta com uma subclasse, ou seja, a quantidade de documentos de patentes

que mencionam o tópico e uma subclasse em particular. É, portanto, a coocorrência de subclasse e o tópico no conjunto de documentos que mencionam ambos. Essa coocorrência (peso) determina a espessura das arestas – quanto mais espessa a aresta, maior a relevância de um tópico na subclasse. Por outro lado, a importância do tópico (nó) é definida pela soma das frequências das arestas que o conecta às suas subclasses.

Figura 37 – Grafo de conhecimento com destaque para algumas subclasses sugeridas



Fonte: elaborado pelo autor (2023)

A representação da Figura 37 enfatiza as subclasses de patentes com maior relevância com base no *ranking* gerado pelo modelo, no exemplo identificado pelas subclasses H04W, H04J e H04B. O examinador, com base na indicação do *ranking* e no grafo de conhecimento, possui elementos que podem auxiliá-lo no processo de tomada de decisão, de tal maneira que a tarefa de classificação no contexto de análise de patentes seja facilitada.

5.4 CONSIDERAÇÕES FINAIS

Para finalizar, tomando-se como base a avaliação do modelo proposto para averiguar seus diversos componentes como solução para o problema apresentado, considera-se que o objetivo foi atingido de forma bem-sucedida. O conjunto de avaliações realizadas tem como

suporte a fundamentação teórica (capítulo 2) e a DSRM (capítulo 3), em que constam todos os elementos (materiais e métodos) necessários para essa etapa. Ou seja, esses capítulos amparam e sustentam a criação e a proposição de cada uma das etapas do modelo, visando atender as lacunas encontradas na literatura.

Mais especificamente, os resultados obtidos no cenário geral e nos cenários específicos oferecem indícios de que a configuração do modelo atual promove resultados importantes para auxiliar na tomada de decisão de examinadores de patentes. O cenário geral efetuou um conjunto expressivo de instanciações, 800 ao todo, variando diferentes configurações no intuito de promover um entendimento mais amplo da interconexão dos elementos que constituem o modelo. Por outro lado, os três cenários específicos objetivaram apresentar, de forma mais detalhada, como a etapa de *ranking* trabalha, promovendo suporte para o entendimento das especificidades do modelo quanto a esse componente.

Ademais, compreendendo que a proposta elaborada apresenta indícios de viabilidade na aplicação operacional, é importante destacar o rigor científico desde a concepção até a condução da pesquisa que resultou no modelo proposto e no conjunto de estratégias utilizadas para a sua avaliação.

Conclui-se que o modelo proposto, baseado em PTMs, estratégias de *ranking* e grafo de conhecimento, configura uma solução efetiva e escalável para apoiar a tarefa de classificação de patentes. De modo geral, demonstrou-se o potencial para prover benefícios práticos aos examinadores de patentes, constituindo uma alternativa viável para essa tarefa.

6 CONCLUSÕES E TRABALHOS FUTUROS

O volume dos dados nas bases de patente cresce de maneira expressiva ano após ano. As patentes depositadas nos escritórios de patentes necessitam passar por um processo de registro. Com isso, aumenta o número de avaliações de patentes realizadas, o que acarreta uma sobrecarga de trabalho aos examinadores. Entre as tarefas executadas por examinadores está a classificação da patente, relacionada à área de Análise de Patentes. Ou seja, trata-se de uma tarefa para rotular as patentes de acordo com a área tecnológica a que pertencem, sendo esse processo muitas vezes desempenhado de forma manual.

A área de Análise de Patentes, mais especificamente a tarefa de classificação de patentes, apresenta diversos desafios para que possa ser efetiva no auxílio aos examinadores. Para manter atualizados os sistemas que auxiliam na classificação de patentes, citam-se a complexidade da linguagem de patentes, a necessidade de processar grandes volumes de patentes, a constante evolução tecnológica e o surgimento de novos campos, as dificuldades em replicar a subjetividade e conhecimento especializado humano, as variações entre sistemas de classificação, a ambiguidade na atribuição de códigos IPC, a necessidade de garantir sugestões de qualidade comparáveis à classificação humana e a demanda e os custos elevados.

Nesse sentido, o presente trabalho procurou atender esses desafios por meio de diferentes métodos e técnicas computacionais e de Engenharia do Conhecimento. Para tal, a revisão integrativa e o referencial teórico forneceram um arcabouço à proposição e ao desenvolvimento de um modelo de recomendação ordenada de subclasses, com características de explicitação de conhecimento e incorporação das decisões efetuadas por examinadores no fluxo do presente trabalho. Mais especificamente, a revisão integrativa e o referencial teórico permitiram a análise de áreas, métodos e técnicas voltados à análise e classificação de patentes, principalmente no âmbito da Inteligência Artificial, do Processamento de Linguagem Natural, do Aprendizado Profundo, dos Modelos Pré-treinados, das Redes Neurais e da Representação de Conhecimento.

Isso posto, consideram-se atendidos os objetivos específicos de identificar métodos e técnicas para sugerir subclasses de maneira ordenada e explicitar o conhecimento latente em bases de patentes, objetivos estes que possibilitam atingir o objetivo geral do trabalho e responder à pergunta de pesquisa.

Como solução para o problema, tem-se ao final deste trabalho um modelo voltado à recomendação ordenada de subclasses de patentes com base em fontes textuais não estruturadas. Para isso, combinam-se a representação vetorial avançada via PTMs, a

recuperação aproximada, a ordenação por relevância e a explicitação visual do conhecimento para apoiar os examinadores na complexa tarefa de classificar patentes.

O modelo converte os textos das patentes em representações vetoriais densas (*embeddings*) por meio de PTMs, em que os *embeddings* são armazenados e indexados em uma base de conhecimento, permitindo consultas aproximadas de maneira eficiente.

Pensando no fluxo do modelo, novos documentos de patentes têm seus *embeddings* comparados aos da base de conhecimento para recomendar, a partir da análise dos documentos de patentes recuperados, as subclasses mais relevantes. Para gerar a recomendação, o modelo efetua consultas vetoriais na base de conhecimento, recuperando patentes próximas e produzindo um *ranking* ordenado das subclasses associadas. O *ranking* é explicitado com o apoio de grafos de conhecimento, elucidando as relações entre subclasses e conceitos.

Com base no *ranking* de subclasses e no grafo de conhecimento, o examinador tem um ferramental para decidir sobre a classificação final mais adequada à patente em análise. Após determinada análise, o resultado da decisão é incorporado à base de dados, permitindo que o modelo seja retroalimentado e mantendo a evolução contínua da base de conhecimento e o aprimoramento das recomendações.

Os resultados do modelo proposto indicam que é possível responder à pergunta de pesquisa. A instanciação do modelo forneceu suporte à tarefa de classificação realizada por examinadores. A avaliação do modelo proposto foi efetivada através de cenários de estudo elaborado com o auxílio do conjunto de dados obtidos na base de patentes do United States Patent and Trademark Office (USPTO®). O conjunto de dados foi preparado para treinamento e teste, em que se utilizou o conteúdo textual para geração de vetores densos (*embeddings*) por diferentes PTMs com arquitetura *transformer*.

Para a recomendação de subclasses, duas estratégias de ordenação (*ranking*) foram propostas: 1) Soma das Ocorrências (SO); e 2) Soma dos *Scores* (SS). Os resultados sugerem que as estratégias de ordenação, que se utilizam de documentos recuperados por meio de similaridade vetorial, são apropriadas para atender o modelo. Para avaliar os *rankings* gerados na etapa de teste, utilizou-se uma métrica de acurácia, não sendo empregadas outras métricas tradicionais para avaliação de classificadores. Ressalta-se que o problema de classificação abordado nesta tese é considerado multissaída (*multi-output*), pois recomenda múltiplas subclasses, e multientrada (*multi-input*), visto que determinada patente pode ter uma ou mais subclasses de entrada.

O modelo obteve melhor desempenho considerando n=50 patentes retornadas e k=5 ou 6 subclasses sugeridas. Valores maiores de n e k apresentaram pouca melhora na acurácia.

Vale ressaltar que valores maiores de *k* impactam positivamente na acurácia, todavia podem dificultar a avaliação, já que o examinador receberá um número maior de subclasses. Apesar disso, pensando em uma implementação do modelo através de um sistema, o examinador poderia configurar o número máximo de subclasses que deseja receber ou visualizar, ou simplesmente ter uma opção que na interface permita a expansão dos resultados. Analisando os resultados, os modelos contextuais (*all-MiniLM-L6-v2*, *all-mpnet-base-v2* e *all-distilroberta-v1*) tiveram desempenho superior ao modelo não contextual (*en_core_web_lg*). Na comparação dos PTMs com as redes neurais, percebe-se que estes também apresentaram desempenho superior em relação às redes neurais MLP, CNN e LSTM, testadas inicialmente na primeira proposta do modelo apresentado na qualificação.

Outro ponto relevante do modelo diz respeito à explicitação do conhecimento, o que foi realizado através do conceito de grafo de conhecimento, permitindo a criação de uma estrutura que vincula tópicos e subclasses. Para tal, utilizou-se a base da DBpedia® objetivando auxiliar no processo de extração de tópicos a partir do texto das patentes. O grafo de conhecimento tem por objetivo explicitar os resultados providos pela recomendação de subclasses, apresentando, de maneira interconectada, seus principais tópicos. Esse tipo de representação permite ao examinador ter mais subsídios sobre o contexto de determinada patente, podendo auxiliá-lo na identificação das subclasses mais adequadas.

Por fim, ressalta-se a dinâmica do modelo ao longo do tempo. Por meio da recomendação de subclasses e do grafo de conhecimento, o examinador possui ferramental importante para decidir as subclasses que melhor representam determinada patente. E uma vez definidas essas subclasses, a patente é vinculada a elas na base de dados. Pensando no fluxo do modelo, ao ocorrer essa atualização a patente passa pela geração de seu *embedding* e de seu grafo de conhecimento, atualizando tanto a base de conhecimento quanto o grafo de conhecimento geral. Dessa forma, novas demandas de classificação utilizarão estruturas que são atualizadas constantemente, tendo assim a expectativa de que recomendações mais assertivas possam ser realizadas com o passar do tempo.

Ademais, considera-se que o objetivo da pesquisa foi alcançado com êxito. O conhecimento do tema teve avanços e discussões a respeito da transformação dos dados, das estratégias de ordenação e da avaliação e dinamicidade do modelo, possibilitando concluir que o modelo proposto e desenvolvido responde à pergunta de pesquisa. Entretanto, limitações são inerentes ao processo, assim como proposições para desenvolvimentos futuros, conforme apresentado na sequência.

6.1 LIMITAÇÕES

A qualidade dos dados de entrada é um fator crucial em modelos de NLP baseados em dados. Esses modelos são sensíveis à qualidade dos dados usados para treinamento, podendo resultar em saídas imprecisas caso contenham muitos ruídos. A quantidade de dados de patentes é outro fator relevante para o treinamento e a extração de relações. Uma base de dados pequena ou muito específica, como a utilizada no cenário de comparação dos PTMs com redes neurais, pode limitar a abrangência do modelo de fazer recomendações assertivas.

Outra limitação inerente ao estudo refere-se à escalabilidade relativa ao grande número de níveis hierárquicos da IPC, necessitando de uma grande quantidade de recursos computacionais para realizar a tarefa de classificação.

Por fim, a construção e a manutenção de um KG requerem esforço computacional e humano, especificamente para domínios dinâmicos e complexos. A representação do conhecimento pode gerar ambiguidade e inconsistências no KG, gerando dificuldade de interpretação.

6.2 PERSPECTIVAS E TRABALHOS FUTUROS

Durante a realização deste trabalho foi possível identificar possíveis pontos de melhoria no modelo proposto que, acredita-se, refletirão nos resultados a serem ofertados aos examinadores. Nesse sentido, vislumbra-se a possibilidade de testes com outras arquiteturas recentes como GPT-4®, Bard® e Llama® para a sumarização dos principais documentos recuperados (*top* 10) e a geração de grafos de conhecimento a partir do texto sumarizado para auxiliar os examinadores.

No tocante à questão temporal, uma análise do impacto do *ranking* considerando documentos mais recentes parece interessante. Isso se justifica pelo fato de que a tecnologia evolui muito rápido, assim como a taxonomia da IPC. Dessa forma, o *ranking* poderia combinar em uma métrica o *score* (relevância semântica das patentes recuperadas) normalizado pelo ano de publicação (temporalidade), possibilitando analisar o impacto temporal nesse *ranking*.

No que tange à representação do conhecimento, torna-se fundamental evoluir a representação do grafo de conhecimento, assim como avançar nas questões técnicas e tecnológicas envolvidas, visto que à medida que os dados crescem e as áreas evoluem, novos tópicos e relações são gerados em larga escala. Uma infraestrutura adequada pode contribuir não somente para o cenário desta tese, mas também para outras tarefas importantes na área de

análise de patentes, como, por exemplo, a previsão de tecnologias por meio da predição de ligações (*link prediction*).

Apesar de não fazer parte do escopo desta tese, durante o seu desenvolvimento testes foram realizados utilizando redes *transformers* para a geração de *embeddings* das imagens das patentes. Os *embeddings* foram combinados com *embeddings* do texto da patente e do próprio título da imagem disponível. Resultados preliminares sugerem a possibilidade de recuperação de documentos para auxiliar na classificação tanto pelo texto quanto pela imagem, fornecendo uma resposta mais precisa a partir de uma demanda de análise requisitada pelo examinador.

Por fim, sugere-se uma solução alternativa ao problema utilizando-se a taxonomia IPC/CPC como fonte de treinamento principal em vez dos documentos de patentes. Essa taxonomia contém hierarquias e relacionamentos entre as diferentes subclasses, assim como uma descrição de cada elemento, o que pode ajudar o modelo a aprender essas associações de forma mais direta. Mais especificamente, em vez de se indexar o documento da patente, seria indexado o conteúdo textual de determinada subclasse ou mesmo o conteúdo de níveis inferiores (subgrupo). Ao realizar a consulta, o *embedding* da patente de entrada seria então comparado com os *embeddings* dos elementos da taxonomia, provendo, de maneira direta, o resultado com as subclasses e o próprio score da consulta realizada em que a ordem de retorno já determinaria o *ranking* das subclasses. Destaca-se ainda que essa abordagem reduz o custo computacional, pois evita que a base de conhecimento cresça rapidamente. Apesar de ser uma possibilidade promissora, a quantidade de texto que descreve cada elemento da estrutura é limitada, o que pode impactar negativamente nos resultados.

O modelo proposto também poderia ser empregado em outros contextos como, por exemplo, sugerir categorias relevantes para novos artigos científicos com base nos artigos previamente classificados em áreas e subáreas do conhecimento. Da mesma forma, o modelo poderia indicar tópicos adequados de conteúdo textual de notícias, *posts* em mídias sociais e processos judiciais. Em todos esses casos, o modelo se beneficiaria de taxonomias ou ontologias predefinidas, com categorias e subcategorias nas quais os novos itens textuais poderiam ser classificados com o apoio das recomendações. A representação vetorial de textos por meio de *embeddings* permite capturar semanticamente o conteúdo dos itens a serem categorizados.

Desse modo, por meio da similaridade vetorial e da análise da frequência/relevância das categorias em itens pré-classificados, o modelo seria capaz de produzir uma sugestão de classes relevantes para novos itens. Portanto, observa-se bom potencial de aplicação da abordagem em diversas tarefas de classificação textual automatizada.

Além das contribuições teóricas e práticas já descritas, este trabalho resultou no seguinte produto:

- O artigo MODELO DE CLASSIFICAÇÃO DE PATENTES BASEADO EM TÉCNICAS DE ENGENHARIA DE CONHECIMENTO, submetido ao XII Congresso Internacional de Conhecimento e Inovação (ciKi). DOI: https://doi.org/10.48090/ciki.v1i1.1254. Disponível em https://proceeding.ciki.ufsc .br/index.php/ciki/article/view/1254.

Dessa forma, além de resumir as contribuições teóricas, a conclusão também destaca os produtos concretos oriundos do trabalho, o que demonstra impacto e permite que outros pesquisadores se beneficiem dos resultados.

REFERÊNCIAS

ABBAS, A.; ZHANG, L.; KHAN, S. U. A literature review on the state-of-the-art in patent analysis. **World Patent Information**, v. 37, p. 3-13, 2014.

ABDELGAWAD, L. *et al*. Optimizing neural networks for patent classification. In: MACHINE LEARNING AND KNOWLEDGE DISCOVERY IN DATABASES: EUROPEAN CONFERENCE, 2019, Würzburg, Germany. **Proceedings** […]. Berlin, Heidelberg: Springer-Verlag, 2019. Sigla do evento: ECML PKDD. p. 16-20, parte III. Disponível em: https://dl.acm.org/doi/abs/10.1007/978-3-030-46133-1_41. Acesso em: 10 jul. 2022.

ABEL, M.; RAMA FIORINI, S. Uma revisão da engenharia do conhecimento: evolução, paradigmas e aplicações. **International Journal of Knowledge Engineering and Management**, v. 2, n. 2, p. 1, 2013.

ABIRAMI, S.; CHITRA, P. Energy-efficient edge based real-time healthcare support system. *In:* RAJ, P.; EVANGELINE, E. (ed.). **Advances in Computers**. [*s. l*.]: Elsevier, 2020. v. 117. p. 339-368.

AGOSTINI, L.; NOSELLA, A.; HOLGERSSON, M. Patent management: the prominent role of strategy and organization. **European Journal of Innovation Management**, v. 26, n. 4, p. 1054-1070, 2023.

AGOSTINI, L.; NOSELLA, A.; TESHOME, M. B. Towards the development of scales to measure patent management. **World Patent Information**, v. 58, p. 101909, 2019.

AKEN, J. E. van. Management research based on the paradigm of the design sciences: the quest for field-tested and grounded technological rules. **Journal of Management Studies**, v. 41, n. 2, p. 219-246, 2004.

ALLAHYARI, M. *et al*. A brief survey of text mining: classification, clustering and extraction techniques. *In*: KDD BIGDAS, 2017, Halifax, Canada. **Proceedings** […]. 2017. v. 13.

ALOM, M. Z. *et al*. A state-of-the-art survey on deep learning theory and architectures. **Electronics**, v. 8, n. 3, 2019. Disponível em: https://www.mdpi.com/2079-9292/8/3/292. Acesso em: 27 jul. 2021.

ALTUNTAS, F. Recent trends, applications and technological evaluation of protective textile with patent analysis. **Kybernetes**, 2023.

ARISTODEMOU, L.; TIETZE, F. The state-of-the-art on Intellectual Property Analytics (IPA): a literature review on artificial intelligence, machine learning and deep learning methods for analysing intellectual property (IP) data. **World Patent Information**, v. 55, p. 37-51, 2018.

AROYEHUN, S. T. *et al*. Leveraging label hierarchy using transfer and multi-task learning: a case study on patent classification. **Neurocomputing**, v. 464, n. 13, p. 421-431, 2021.

BABIĆ, K.; MARTINČIĆ-IPŠIĆ, S.; MEŠTROVIĆ, A. Survey of neural text representation models. **Information**, v. 11, n. 11, 2020. Disponível em: https://www.mdpi.com/2078-2489/11/11/511/htm. Acesso em: 25 nov. 2021

BAGLIERI, D.; CESARONI, F. Capturing the real value of patent analysis for R&D strategies. **Technology Analysis and Strategic Management**, v. 25, n. 8, p. 971-986, 2013.

BAI, J.; SHIM, I.; PARK, S. MEXN: multi-stage extraction network for patent document classification. **Applied Sciences**, v. 10, n. 18, p. 6229, 2020.

BASKERVILLE, R.; DULIPOVICI, A. The theoretical foundations of knowledge management. **Knowledge Management Research & Practice**, v. 4, n. 2, p. 83-105, 2006.

BAX, M. P. Design science: filosofia da pesquisa em ciência da informação e tecnologia. **Ciencia da Informação**, v. 42, n. 2, p. 298-312, 2013.

BECATTINI, N. *et al*. ARIZ85 and patent-driven knowledge support. **Procedia Engineering**, v. 131, p. 291-302, 2015.

BONINO, D.; CIARAMELLA, A.; CORNO, F. Review of the state-of-the-art in patent information and forthcoming evolutions in intelligent patent informatics. **World Patent Information**, v. 32, n. 1, p. 30-38, 2010.

BOTELHO, L. L. R.; CUNHA, C. C. DE A.; MACEDO, M. O método da revisão integrativa nos estudos organizacionais. **Gestão e Sociedade**, v. 5, n. 11, p. 121, 2011.

BRACHMAN, R. J.; LEVESQUE, H. J. **In praise of knowledge representation and reasoning**. San Francisco: Elsevier, 2004.

BROWN, T. B. *et al*. Language models are few-shot learners. *In:* CONFERENCE ON NEURAL INFORMATION PROCESSING SYSTEMS, 34., 2020, Vancouver, Canada. **Proceedings** […]. Vancouver, Canada: Neural Information Processing Systems Foundation, 2020. Sigla do evento: NeurIPS 2020. Disponível em: https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf. Acesso em: 8 dez. 2021.

BUSCALDI, D. *et al*. Answering questions with an n-gram based passage retrieval engine. **Journal of Intelligent Information Systems**, v. 34, n. 2, p. 113-134, 2010.

CAMACHO-COLLADOS, J.; PILEHVAR, M. T. From word to sense embeddings: a survey on vector representations of meaning. **Journal of Artificial Intelligence Research**, v. 63, p. 743-788, 2018.

CAPES. **Documento de Área 45: Interdisciplinar**. Disponível em: https://www.gov.br/capes/pt-br/acesso-a-informacao/acoes-e-programas/avaliacao/sobre-a-avaliacao/areas-avaliacao/sobre-as-areas-de-avaliacao/colegio-de-ciencias-exatas-tecnologicas-e-multidisciplinar/multidisciplinar/interdisciplinar. Acesso em: 18 nov. 2019.

CASSIDY, C. Parameter tuning Naïve Bayes for automatic patent classification. **World Patent Information**, v. 61, p. 101968, 2020.

CEN. **European Guide to good Practice in Knowledge Management - Part 1: Knowledge Management Framework**. Brussels, 2004.

CHAO, M.-H. *et al*. Technology mining for intelligent chatbot development. *In*: NEWNES *et al*. (ed.). **Transdisciplinary engineering for resilience**: responding to system disruptions. 2021. (Série Advances in Transdisciplinary Engineering, v. 16, p. 123-132). Disponível em: https://ebooks.iospress.nl/doi/10.3233/ATDE210090. Acesso em: 29 nov. 2021.

CHEN, X.; JIA, S.; XIANG, Y. A review: knowledge reasoning over knowledge graph. **Expert Systems with Applications**, v. 141, p. 112948, 2020.

CHEN, Z. *et al*. Knowledge graph completion: a review. **IEEE Access**, v. 8, p. 192435-192456, 2020.

CHOI, S. *et al*. Deep learning for patent landscaping using transformer and graph embedding. **Technological Forecasting and Social Change**, v. 175, p. 121413, 2022.

CHOI, S. Y.; KIM, S. H. Knowledge acquisition and representation for high-performance building design: a review for defining requirements for developing a design expert system. **Sustainability**, v. 13, n. 9, p. 4640, 2021. Disponível em: https://www.mdpi.com/2071-1050/13/9/4640/htm. Acesso em: 30 nov. 2021.

CHOI, Y.; HONG, S. Qualitative and quantitative analysis of patent data in nanomedicine for bridging the gap between research activities and practical applications. **World Patent Information**, v. 60, p. 101943, 2020.

CHOLLET, F. **Deep learninf with Python**. Shelter Island, NY: Manning, 2017.

CHUNG, P.; SOHN, S. Y. Early detection of valuable patents using a deep learning model: case of semiconductor industry. **Technological Forecasting and Social Change**, v. 158, p. 120146, 2020.

CRESWELL, J. W. **Projeto de pesquisa:** métodos qualitativo, quantitativo e misto. 2. ed. Porto Alegre: Artmed, 2010.

CUPANI, A. La peculiaridad del conocimiento tecnológico. **Scientia Studia**. v. 4, n. 3, p. 353-371, 2006.

DATA SCIENCE ACADEMY. **Deep Learning Book**. Disponível em: https://www.deeplearningbook.com.br/arquitetura-de-redes-neurais-long-short-term-memory. Acesso em: 17 fev. 2022.

DE CLERCQ, D. *et al*. Multi-label classification and interactive NLP-based visualization of electric vehicle patent data. **World Patent Information**, v. 58, p. 101903, 2019.

DEL CARPIO RAMOS, H. A.; DEL CARPIO RAMOS, P. A.; GARCÍA-PEÑALVO, F. J. Technological research methodology to manage organizational change. *In*: INTERNATIONAL CONFERENCE ON TECHNOLOGICAL ECOSYSTEMS FOR ENHANCING MULTICULTURALITY, 7., 2019. **Proceedings** […]. 2019. Sigla do evento: TEEM´19. p. 168-176. Disponível em: https://doi.org/10.1145/3362789.3362890. Acesso em: 28 dez. 2021.

DENG, W.; HUANG, X.; ZHU, P. Facilitating technology transfer by patent knowledge graph. *In*: HAWAII INTERNATIONAL CONFERENCE ON SYSTEM SCIENCES, 52., 2019. **Proceedings** […]. 2019. Disponível em: http://hdl.handle.net/10125/59566. Acesso em: 16 jan. 2020

DEVLIN, J. *et al*. BERT: Pre-training of deep bidirectional transformers for language understanding. *In*: ANNUAL CONFERENCE OF THE NORTH AMERICAN CHAPTER OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, 2019, Minneapolis, Minnesota. **Proceedings** […]. ACL, 2019. p. 4171-4186. Disponível em: https://aclanthology.org/N19-1423.pdf . Acesso em: 26 jan. 2021.

DRESCH, A.; LACERDA, D. P.; ANTUNES JR., J. A. V. **Design science research**: a method for science and technology advancement. [*s. l.*]: Springer, 2015.

EGC. Programa de Pós-Graduação em Engenharia e Gestão do Conecimento (PPGEGC). **Áreas de concentração**. Disponível em: https://ppgegc.paginas.ufsc.br/areas-de-concentracao. Acesso em: 15 jul. 2019.

EHRLINGER, L.; WÖß, W. Towards a definition of knowledge graphs. *In:* INTERNATIONAL CONFERENCE ON SEMANTIC SYSTEMS, 12., 2016, Leipzig, Germany. **Proceedings** […]. New York, NY: ACM, 2016. Sigla do evento: SEMANTiCS 2016.

EKMAN, M. **Learning deep learning:** theory and practice of neural networks, computer vision, NLP, and transformers using TensorFlow. Addison-Wesley, 2021. Disponível em: https://www.oreilly.com/library/view/learning-deep-learning/9780137470198. Acesso em: 21 out. 2022

ERNST, H. Patent information for strategic technology management. **World Patent Information**, v. 25, n. 3, p. 233-242, 2003.

EVANGELISTA, A. *et al*. Unveiling the technological trends of augmented reality: a patent analysis. **Computers in Industry**, v. 118, p. 103221, 1 jun. 2020.

FALL, C. J.; BENZINEB, K. Literature survey: issues to be considered in the automatic classification of patents. **World Intellectual Property Organization**, p. 1-64, 2002.

FRERICH, K. *et al*. On the potential of taxonomic graphs to improve applicability and performance for the classification of biomedical patents. **Applied Sciences**, v. 11, n. 2, p. 690, 2021.

GASSMANN, O.; BADER, M. A.; THOMPSON, M. J. **Patent management:** protecting intellectual property and innovation. [*s. l.*]: Springer, 2021.

GERHARDT, T.; SILVEIRA, D. **Métodos de pesquisa**. Porto Alegre: Editora da UFRGS, 2009.

GETOOR, L.; MACHANAVAJJHALA, A. Entity resolution: theory, practice & open challenges. **Proceedings of the VLDB Endowment**, v. 5, n. 12, p. 2018-2019, 2012.

GEUM, Y.; KIM, M. How to identify promising chances for technological innovation: keygraph-based patent analysis. **Advanced Engineering Informatics**, v. 46, p. 101155, 1 out. 2020.

GIL, A. C. **Métodos e técnicas de pesquisa social**. 6. ed. São Paulo: Atlas, 2008.

GIRTHANA, K.; SWAMYNATHAN, S. Query-Oriented Patent Document Summarization System (QPSS). *In:* PANT, M. *et al*. (ed.). **Soft computing**: theories and applications. Advances in intelligent systems and computing. Singapore: Springer, 2020. v. 1053, p. 237-246. DOI: https://doi.org/10.1007/978-981-15-0751-9_22

GOMEZ, J. C.; MOENS, M.-F. A survey of automated hierarchical classification of patents. *In*: PALTOGLOU, G.; LOIZIDES, F.; HANSEN, P. (ed.). Professional **search in the modern world**. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2014. v. 8830, p. 215-249. DOI: doi.org/10.1007/978-3-319-12511-4_11.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. **Deep learning**. Cambridge: MIT Press, 2016. Disponível em: https://books.google.com.br/books?hl=pt-BR&lr=&id=omivDQAAQBAJ&oi=fnd&pg=PR5&ots=MNO5duqGNV&sig=bTk-Sdv1xlmz2-yol7437DmGyfk&redir_esc=y#v=onepage&q&f=false. Acesso em: 27 jul. 2021

GRAWE, M. F.; MARTINS, C. A.; BONFANTE, A. G. Automated patent classification using word embedding. *In*: INTERNATIONAL CONFERENCE ON MACHINE LEARNING AND APPLICATIONS, 16., 2017, Cancun, Mexico. **Proceedings** […]. Cancun, Mexico: IEEE, 2017. Sigla do evento: ICMLA. Disponível em: http://ieeexplore.ieee.org/document/8260665. Acesso em: 11 ago. 2021. p. 408-411.

GREGÓRIO, J. *et al*. The role of Design Science Research Methodology in developing pharmacy eHealth services. **Research in Social and Administrative Pharmacy**, v. 17, n. 12, p. 2089-2096, 2021.

GROVER, A.; LESKOVEC, J. Node2vec: scalable feature learning for networks. *In*: ACM SIGKDD INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING, 22., 2016, New York. **Proceedings** […]. New York: ACM, 2016. p. 855-864. Disponível em: http://dx.doi.org/10.1145/2939672.2939754. Acesso em: 8 dez. 2021.

HAGHIGHIAN ROUDSARI, A. *et al*. PatentNet: multi-label classification of patent documents using deep learning based language understanding. **Scientometrics**, v. 127, n. 1, p. 207-231, 2022.

HEVNER *et al*. Design science in information systems research. **MIS Quarterly**, v. 28, n. 1, p. 75, 2004.

HEVNER, A. R. Design research in food science: keynote introduction. *In*: INTERNATIONAL CONFERENCE ON DATA ENGINEERING WORKSHOPS, 36., 2020, Dallas, Texas. **Proceedings** […]. IEEE, 2020. Sigla do evento: ICDEW. Disponível em: https://ieeexplore.ieee.org/document/9094108. Acesso em: 7 jan. 2022.

HONG, M.; WANG, H. Research on customer opinion summarization using topic mining and deep neural network. **Mathematics and Computers in Simulation**, v. 185, p. 88-114, 2021.

HU, J. *et al*. A hierarchical feature extraction model for multi-label mechanical patent classification. **Sustainability**, v. 10, n. 1, p. 219, 2018.

HUANG, J. Y.; TAN, K. W. An extension-based classification system of cloud computing patents. **International Journal of Information Technology And Decision Making**, v. 19, n. 4, p. 1149-1172, 2020.

HUANG, L. *et al*. A semi-supervised learning framework for TRIZ-based Chinese patent classification. *In*: INTERNATIONAL CONFERENCE ON COMPUTING AND ARTIFICIAL INTELLIGENCE, 6., 2020, New York. **Proceedings** […]. ACM, 2020. p. 46-50. Disponível em: https://doi.org/10.1145/3404555.3404600. Acesso em: 20 dez. 2021.

INPI. **A propriedade intelectual e o comércio exterior**: conhecendo oportunidades para seu negócio. Rio de Janeiro: INPI, 2018. Disponível em: https://www.gov.br/inpi/pt-br/composicao/arquivos/pi_e_comercio_exterior_inpi_e_apex.pdf. Acesso em: 20 dez. 2021.

JAFERY, W. A. Z. W. C. *et al*. Classification of patents according to industry 4.0 pillars using machine learning algorithms. *In*: INTERNATIONAL CONFERENCE ON RESEARCH AND INNOVATION IN INFORMATION SYSTEMS, 6., 2019, Johor Bahru, Malaysia, 2019. **Proceedings** […]. IEEE, 2019. Sigla do evento: ICRIIS. p. 1-6. Disponível em: https://ieeexplore.ieee.org/document/9073669. Acesso em: 23 dez. 2020.

JANG, H.; YOON, B. TechWordNet: development of semantic relation for technology information analysis using F-term and natural language processing. **Information Processing & Management**, v. 58, n. 6, p. 102752, 2021.

JÄRVINEN, P. Action research is similar to design science. **Quality and Quantity**, v. 41, n. 1, p. 37-54, 2007.

JI, S. *et al*. A survey on knowledge graphs: representation, acquisition and applications. **IEEE Transactions on Neural Networks and Learning Systems**, v. 33, n. 2, p. 494-514, 2020.

JIANG, S. *et al*. Deep learning for technical document classification. **IEEE Transactions on Engineering Management**, p. 1-17, 2022.

JURAFSKY, D.; MARTIN, J. **Speech and language processing**. An introduction to natural language processing, computational linguistics, and speech recognition. 3. ed. [*s. l.*: *s. n.*], 2023.

KALYAN, K. S.; SANGEETHA, S. SECNLP: a survey of embeddings in clinical natural language processing. **Journal of Biomedical Informatics**, v. 101, p. 103323, 2020.

KEJRIWAL, M. Knowledge graphs and covid-19: opportunities, challenges, and implementation. **Harvard Data Science Review**, 2020.

KEJRIWAL, M.; MIRANKER, D. P. An unsupervised instance matcher for schema-free RDF data. **Journal of Web Semantics**, v. 35, p. 102-123, 2015.

KIM, J. *et al*. Patent document clustering with deep embeddings. **Scientometrics**, v. 123, n. 2, p. 563-577, 2020.

KIM, J.-H. *et al*. Accelerating large-scale graph-based nearest neighbor search on a computational storage platform. **IEEE Transactions on Computers**, v. 72, n. 1, p. 278-290, 2023.

KIM, S.; PARK, I.; YOON, B. Sao2vec: development of an algorithm for embedding the subject-action-object (SAO) structure using Doc2Vec. **PLoS ONE**, v. 15, n. 2, p. e0227930, 2020.

KITCHENHAM, B.; CHARTERS, S. **Guidelines for performing systematic literature reviews in software engineering**. Technical report. 2007. v. 2.

HINKELMANN, K.; LAUX, A. (ed.). Knowledge Representation Techniques**.** *In:* DFKI-WORKSHOP, 1993, Kaiserslautern. **Proceedings** […]. Kaiserslautern, 8 jul. 1993.

KO, N. *et al*. A transferability evaluation model for intellectual property. **Computers & Industrial Engineering**, v. 131, p. 344-355, 2019.

KORDE, V.; MAHENDER, C. N. Text classification and classifiers: a survey. **International Journal of Artificial Intelligence & Applications**, v. 3, n. 2, p. 85-99, 2012.

KRESTEL, R. *et al*. A survey on deep learning for patent analysis. **World Patent Information**, v. 65, p. 102035, 2021.

KRÖTZSCH, M.; SIMANČÍK, F.; HORROCKS, I. A description logic primer. **Perspectives on Ontology Learning**, v. 18, 2012.

KUMAR, B. S.; RAVI, V. A survey of the applications of text mining in financial domain. **Knowledge-Based Systems**, v. 114, p. 128-147, 2016.

LAN, Z. *et al*. ALBERT: a lite BERT for self-supervised learning of language representations. *In*: INTERNATIONAL CONFERENCE ON LEARNING REPRESENTATIONS. **Proceedings** […]. 26 set. 2020. Sigla do evento: ICLR 2020. Disponível em: https://arxiv.org/abs/1909.11942v6. Acesso em: 3 ago. 2023

LAURIOLA, I.; LAVELLI, A.; AIOLLI, F. An introduction to deep learning in natural language processing: models, techniques, and tools. **Neurocomputing**, v. 470, p. 443-456, 2022.

LI, S. *et al*. DeepPatent: patent classification with convolutional neural networks and word embedding. **Scientometrics**, v. 117, n. 2, p. 721-744, 2018.

LI, W. *et al*. Approximate nearest neighbor search on high dimensional data - experiments, analyses, and improvement. **IEEE Transactions on Knowledge and Data Engineering**, v. 32, n. 8, p. 1475-1488, 2020.

LIN, Y. *et al*. Knowledge representation learning: a quantitative review. **ArXiv**, v. abs/1812.10901, 2018.

LIST, J. Current trends and future directions for IP information research. **World Patent Information**, v. 52, p. A1-A2, 2018.

LIU, L. *et al*. A new function-based patent knowledge retrieval tool for conceptual design of innovative products. **Computers in Industry**, v. 115, p. 103154, 2020.

LIU, X. *et al*. A patent recommendation algorithm based on topic classification and semantic similarity. *In*: INTERNATIONAL CONFERENCE ON WIRELESS COMMUNICATIONS AND SMART GRID, Hangzhou, China, 2021. **Proceedings** […]. IEEE, 2021. Sigla do evento: ICWCSG. p. 289-292. Disponível em: https://ieeexplore.ieee.org/document/9616588. Acesso em: 10 jul. 2021.

LIU, Y. *et al*. RoBERTa: a robustly optimized BERT pretraining approach. **ArXiv**, v. 1907.11692v1, 2019.

LO, C. C.; CHO, H. C.; WANG, P. W. Global R&D collaboration in the development of nanotechnology: the impact of R&D collaboration patterns on patent quality. **Sustainability**, v. 12, n. 15, p. 1-12, 2020.

LO, H.-C.; CHU, J.-M. Pre-trained transformer-based classification for automated patentability examination. *In*: ASIA-PACIFIC CONFERENCE ON COMPUTER SCIENCE AND DATA ENGINEERING, 2021, Brisbane, Australia. **Proceedings** […]. IEEE, 2021. Sigla do evento: CSDE. p. 1-5. Disponível em: https://ieeexplore.ieee.org/document/9718474. Acesso em: 20 abr. 2022.

LU, H. *et al*. A patent text classification model based on multivariate neural network fusion. *In*: INTERNATIONAL CONFERENCE ON SOFT COMPUTING & MACHINE INTELLIGENCE, 6., 2019, Johannesburg, South Africa. **Proceedings** […]. IEEE, 2019. Sigla do evento: ISCMI. Disponível em: https://ieeexplore.ieee.org/document/9004335. Acesso em: 11 ago. 2021.

MALKOV, Y. *et al*. Approximate nearest neighbor algorithm based on navigable small world graphs. **Information Systems**, v. 45, p. 61-68, 2014.

MAO, Q.; TSANG, I. W. H.; GAO, S. Objective-guided image annotation. **IEEE Transactions on Image Processing**, v. 22, n. 4, p. 1585-1597, 2013.

MARCONI, M. de A.; LAKATOS, E. M. **Fundamentos de metodologia científica**. 5. ed. São Paulo: Atlas, 2003.

MEGURO, K.; OSABE, Y. Lost in patent classification. **World Patent Information**, v. 57, p. 70-76, 2019.

MEIRELES, M. R. G.; FERRARO, G.; GEVA, S. Classification and information management for patent collections: a literature review and some research questions. **Information Research**, v. 21, n. 1, 2016. Disponível em: http://www.informationr.net/ir/21-1/paper705.html#.X_33wuhKhPZ. Acesso em: 12 jan. 2021.

MIKOLOV, T. *et al*. Distributed representations of words and phrases and their compositionality. *In*: CONFERENCE ON NEURAL INFORMATION PROCESSING SYSTEMS, 27., 2013a, Lake Tahoe, Nevada, Estados Unidos. **Proceedings** […]. NeurIPS, 2013a. Sigla do evento: NeurIPS. Disponível em: http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-com.pdf. Acesso em: 7 dez. 2021a.

MIKOLOV, T. *et al*. Efficient estimation of word representations in vector space. *In*: INTERNATIONAL CONFERENCE ON LEARNING REPRESENTATIONS, 1., 2023b. **Proceedings** […]. NeurIPS, 2013b. p. 1-12. Disponível em: http://arxiv.org/abs/1301.3781. Acesso em: 7 dez. 2021.

MIN, H. Power patent classification method based on deep neural network. **Journal of Physics**: Conference Series, v. 1848, n. 1, p. 012048, 2021.

MISHRA, M. *et al*. Deep learning in electrical utility industry: a comprehensive review of a decade of research. **Engineering Applications of Artificial Intelligence**, v. 96, p. 104000, 2020.

MOEHRLE, M. G. *et al*. Patinformatics as a business process: a guideline through patent research tasks and tools. **World Patent Information**, v. 32, n. 4, p. 291-299, 2010. Disponível em: https://linkinghub.elsevier.com/retrieve/pii/S0172219009001380. Acesso em: 18 maio 2021.

MOEHRLE, M. G.; WALTER, L.; WUSTMANS, M. Designing the 7D patent management maturity model; a capability based approach. **World Patent Information**, v. 50, p. 27-33, 2017.

NADEAU, D.; SEKINE, S. A survey of named entity recognition and classification. **Lingvisticae Investigationes**, v. 30, n. 1, p. 3-26, 2007.

NAIK, D. A.; BRUNDA, C. J.; SEMA, S. A feasible dashboard to predict patent mining using classification algorithms. **Procedia Computer Science**, v. 167, p. 2011-2021, 2020.

NAZÁRIO, D. C.; DANTAS, M. A. R.; TODESCO, J. L. Knowledge engineering: survey of methodologies, techniques and tools. **IEEE Latin America Transactions**, v. 12, n. 8, p. 1553-1559, 2014.

NOH, H.; LEE, S. What constitutes a promising technology in the era of open innovation? An investigation of patent potential from multiple perspectives. **Technological Forecasting and Social Change**, v. 157, p. 120046, ago. 2020.

OTTER, D. W.; MEDINA, J. R.; KALITA, J. K. A survey of the usages of deep learning for natural language processing. **IEEE Transactions on Neural Networks and Learning Systems**, v. 32, n. 2, p. 604-624, 2021.

PACHECO, R. C. S. Coprodução em ciência, tecnologia e inovação: fundamentos e visões. *In*: PEDRO, J. M.; FREIRE, P. S. (org.). **Interdisciplinaridade**: universidade e inovação social e tecnológica. Curitiba: CRV, 2016. p. 21-26.

PACHECO, R. C. **Dados e governo abertos na sociedade do conhecimento**. Florianópolis, 19 nov. 2014. 80 slides. Disponível em: http://www.inf.ufsc.br/~jose.todesco/LODBrasil/Abertura/DadosEGovernoAbertoNaSocConh.pdf.

PARK, H.; REE, J. J.; KIM, K. An SAO-based approach to patent evaluation using TRIZ evolution trends. *In:* INTERNATIONAL CONFERENCE ON MANAGEMENT OF INNOVATION & TECHNOLOGY, 2012, Bali, Indonesia. **Proceedings** […]. IEEE, 2012. Sigla do evento: ICMIT. p. 594-598. Disponível em: http://ieeexplore.ieee.org/document/6225873. Acesso em: 2 ago. 2023.

PEFFERS, K. *et al*. A design science research methodology for information systems research. **Journal of Management Information Systems**, v. 24, n. 3, p. 45-77, 2007.

PETERS, M. E. *et al*. Deep contextualized word representations. *In*: CONFERENCE OF THE NORTH AMERICAN CHAPTER OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS: HUMAN LANGUAGE TECHNOLOGIES, 2018, New Orleans, Louisiana. **Proceedings** […]. ACL, 2018. v. 1, p. 2227-2237. Disponível em: https://aclanthology.org/N18-1202. Acesso em: 8 dez. 2021.

PIROI, F.; LUPU, M.; HANBURY, A. Overview of CLEF-IP 2013 lab: information retrieval in the patent domain. *In*: FORNER, P. *et al.* (ed.). **Information access evaluation**: multilinguality, multimodality, and visualization. CLEF 2013. Lecture notes in Computer Science. Berlin, Heidelberg: Springer, 2013. v. 138, p. 232-249. Disponível em: https://link.springer.com/chapter/10.1007/978-3-642-40802-1_25. Acesso em: 12 nov. 2021.

PLOSKAS, N. *et al*. Evaluating and ranking patents with multiple criteria: how many criteria are required to find the most promising patents? **Computers and Chemical Engineering**, v. 123, p. 317-330, 2019.

PONTA, L. *et al*. Innovation capability of firms: a big data approach with patents. *In*: ONETO, L. *et al*. (ed.). **Recent advances in big data and deep learning**. Proceedings of the International Neural Networks Society, Springer, Cham, 2020. Disponível em: https://doi.org/10.1007/978-3-030-16841-4_18. Acesso em: 12 nov. 2021. v. 1, p. 169-179.

PUJARA, J. *et al*. Knowledge graph identification. *In*: ALANI, H. *et al*. **The Semantic Web – ISWC 2013**. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2013. v. 8218. Disponível em: https://link.springer.com/chapter/10.1007/978-3-642-41335-3_34. Acesso em: 22 mar. 2021.

RADFORD, A. *et al*. Improving language understanding by generative pre-training. **OpenAI.com**, p. 1-12, 2018.

RAFFEL, C. *et al*. Exploring the limits of transfer learning with a unified text-to-text transformer. **Journal of Machine Learning Research**, v. 21, p. 1-67, 2020.

REIMERS, N.; GUREVYCH, I. Sentence-BERT: sentence embeddings using siamese BERT-networks. *In*: CONFERENCE ON EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING AND 9TH INTERNATIONAL JOINT CONFERENCE ON NATURAL LANGUAGE PROCESSING, 2019, Hong Kong, China. **Proceedings** […]. ACL, 2019. p. 3982-3992. Disponível em: https://aclanthology.org/D19-1410.pdf. Acesso em: 26 jun. 2023

RENGA BASHYAM, K. G.; VADHIYAR, S. Fast scalable approximate nearest neighbor search for high-dimensional data. *In*: INTERNATIONAL CONFERENCE ON CLUSTER COMPUTING, 2020, Kobe, Japan. **Proceedings** […]. IEEE, 2020. p. 294-302. Disponível em: https://ieeexplore.ieee.org/document/9229578. Acesso em: 20 jul. 2023.

RISCH, J.; KRESTEL, R. Domain-specific word embeddings for patent classification. **Data Technologies and Applications**, v. 53, n. 1, p. 108-122, 2019.

RISCH, J.; KRESTEL, R. Learning patent speak: investigating domain-specific word embeddings. *In*: INTERNATIONAL CONFERENCE ON DIGITAL INFORMATION MANAGEMENT, 13., 2018, Berlin, Germany. **Proceedings** […]. IEEE, 2018. Sigla do evento: ICDIM. p. 63-68. Disponível em: https://ieeexplore.ieee.org/document/8846972. Acesso em: 8 abr. 2021.

ROTHMAN, D. **Transformers for natural language processing**: build innovative deep neural network architectures for NLP with Python, PyTorch, TensorFlow, BERT, RoBERTa, & more. Packt, 2021. Disponível em: https://www.packtpub.com/product/transformers-for-natural-language-processing/9781800565791. Acesso em: 21 out. 2022

ROUDSARI, A. H. *et al*. Comparison and analysis of embedding methods for patent documents. *In*: INTERNATIONAL CONFERENCE ON BIG DATA AND SMART COMPUTING, 2021, Jeju Island, Korea (South). **Proceedings** […]. IEEE, 2021. Sigla do evento: BigComp 2021. p. 152-155. Disponível em: https://ieeexplore.ieee.org/document/9373099. Acesso em: 8 abr. 2021.

ROUDSARI, A. H. *et al*. Multi-label patent classification using attention-aware deep learning model. *In*: INTERNATIONAL CONFERENCE ON BIG DATA AND SMART COMPUTING, 2020, Busan, Korea (South). **Proceedings** […]. IEEE, 2020. Sigla do evento: BigComp 2020. p. 558-559. Disponível em: https://ieeexplore.ieee.org/document/9070766. Acesso em: 8 abr. 2021.

RUIJIE, Z. *et al*. Patent text modeling strategy and its classification based on structural features. **World Patent Information**, v. 67, p. 102084, 2021.

RUSSELL, S. J.; NORVIG, P.; CHANG, M. **Artificial intelligence**: a modern approach. 4. ed. United Kingdom: Pearson Education, 2021.

SAMPIERI, R. H.; COLLADO, C. F.; LUCIO, P. B. **Metodologia de pesquisa**. 5. ed. Porto Alegre: Penso, 2013.

SARKER, I. H. Deep learning: a comprehensive overview on techniques, taxonomy, applications and research directions. **SN Computer Science**, v. 2, n. 6, p. 1-20, 2021a.

SARKER, I. H. Machine learning: algorithms, real-world applications and research directions. **SN Computer Science**, v. 2, n. 3, p. 1-21, 2021b.

SARKER, I. H.; FURHAD, M. H.; NOWROZY, R. AI-Driven Cybersecurity: an overview, security intelligence modeling and research directions. **SN Computer Science**, v. 2, n. 3, p. 1-18, 2021.

SEVERINO, A. J. **Metodologia do trabalho científico**. São Paulo: Cortez, 2013.

SHAHID, M. *et al*. Automatic patents classification using supervised machine learning. Advances in intelligent systems and computing. *In*: INTERNATIONAL CONFERENCE ON SOFT COMPUTING AND DATA MINING, 4., 2020. **Proceedings** […]. Springer, 2020. Disponível em: https://link.springer.com/chapter/10.1007/978-3-030-36056-6_29. Acesso em: 8 jan. 2021.

SHALABY, W.; ZADROZNY, W. Patent retrieval: a literature review. **Knowledge and Information Systems**, v. 22, n. 3, p. 1-30, 2019.

SIMON, H. A. **The sciences of the artificial**. 3. ed. Cambridge: MIT Press, 1996.

SINGHAL, A. Introducing the knowledge graph: things, not strings. **Official Google Blog**, 16 maio 2012.

SOFEAN, M. Deep learning based pipeline with multichannel inputs for patent classification. **World Patent Information**, v. 66, p. 102060, 2021.

SOMAYA, D. How patent strategy affects the timing and method of patent litigation resolution. **Advances in Strategic Management**, v. 34, p. 471-504, 2016.

SORANZO, B.; NOSELLA, A.; FILIPPINI, R. Managing firm patents: a bibliometric investigation into the state of the art. **Journal of Engineering and Technology Management - JET-M**, v. 42, p. 15-30, 2016.

SORCE, S. *et al*. A novel visual interface to foster innovation in mechanical engineering and protect from patent infringement. **Journal of Physics:** Conference Series, v. 1004, n. 1, p. 012024, 2018.

SPACY. **English**. spaCy Models Documentation. Disponível em: https://spacy.io/models/en. Acesso em: 4 ago. 2023.

TEOFILI, T. **Deep learning for search**. **Deep Learning using R**, v. 1, n. MEAP, 2018.

TEOFILI, T. **Deep learning for search**. Manning, 2019.

TRAPPEY, A. J. C. *et al*. A patent quality analysis for innovative technology and product development. **Advanced Engineering Informatics**, v. 26, n. 1, p. 26-34, 2012.

TRAPPEY, A.; TRAPPEY, C. V.; HSIEH, A. An intelligent patent recommender adopting machine learning approach for natural language processing: a case study for smart machinery technology mining. **Technological Forecasting and Social Change**, v. 164, p. 120511, 2021.

TSENG, Y.-H.; LIN, C.-J.; LIN, Y.-I. Text mining techniques for patent analysis. **Information Processing & Management**, v. 43, n. 5, p. 1216-1247, 2007.

USPTO. **Become a patent examiner**. 20 fev. 2020. Disponível em: https://www.uspto.gov/jobs/become-patent-examiner. Acesso em: 20 out. 2021.

VASWANI, A. *et al*. Attention is all you need. *In*: CONFERENCE ON NEURAL INFORMATION PROCESSING SYSTEMS, 31., 2017, Long Beach, CA, USA. **Proceedings** […], 2017. Sigla do evento: NIPS 2017. p. 5999-6009.

WANG, D. *et al*. Research on optimization of big data construction engineering quality management based on RNN-LSTM. **Complexity**, v. 2018, Article ID 9691868, 16 pages, 2018.

WANG, Q. *et al*. Knowledge graph embedding: a survey of approaches and applications. **IEEE Transactions on Knowledge and Data Engineering**, v. 29, n. 12, p. 2724-2743, 2017. Disponível em: http://ieeexplore.ieee.org/document/8047276. Acesso em: 6 abr. 2021.

WANG, Y. *et al*. A CNN-based feature extraction scheme for patent analysis. *In*: INTERNATIONAL CONFERENCE ON COMPUTER AND COMMUNICATIONS, 4., 2018, Chengdu, China. **Proceedings** […]. IEEE, 2018. Sigla do evento: ICCC. p. 2387-2391. Disponível em: https://ieeexplore.ieee.org/document/8780690. Acesso em: 6 jul. 2021.

WIPO. **International Patent Classification (IPC)**. CH-1211 Geneva 20, Switzerland: WIPO, 2022a.

WIPO. **WIPO - Patents**. Disponível em: https://www.wipo.int/patents/en. Acesso em: 11 abr. 2021b.

WIPO. WIPO Technology Trends 2019: Artificial Intelligence. 2019. p. 158.

WIPO. **World Intellectual Property Indicators 2020**. 2020. Geneva: [*s. n*.]. Disponível em: https://www.wipo.int/publications/en/details.jsp?id=4526&plang=EN. Acesso em: 10 jul. 2022.

WIPO. **World Intellectual Property Indicators 2021.**Geneva: World Intellectual Property Organization, 2021a. Disponível em: https://tind.wipo.int/record/44461/files/wipo_pub_941_2021.pdf CN - K1401. Acesso em: 10 jul. 2022.

WIPO. **World Intellectual Property Indicators 2022**. Geneva: World Intellectual Property Organization, 2022a.

WU, J.-L. *et al*. A patent quality analysis and classification system using self-organizing maps with support vector machine. **Applied Soft Computing**, v. 41, p. 305-316, 2016.

WU, J.-L. Patent quality classification system using the feature extractor of deep recurrent neural network. *In*: INTERNATIONAL CONFERENCE ON BIG DATA AND SMART COMPUTING, 2019, Kyoto, Japan. **Proceedings** […]. IEEE, 2019. Sigla do evento: BigComp. Disponível em: https://ieeexplore.ieee.org/document/8679141. Acesso em: 19 out. 2020.

XIAO, L.; WANG, G.; LIU, Y. Patent text classification based on naive bayesian method. *In*: INTERNATIONAL SYMPOSIUM ON COMPUTATIONAL INTELLIGENCE AND DESIGN, 11., 2018, Hangzhou, China. **Proceedings** […]. IEEE, 2018. Sigla do evento: ISCID. p. 57-60. Disponível em: https://ieeexplore.ieee.org/document/8695601. Acesso em: 23 dez. 2020.

XIAO, L.; WANG, G.; ZUO, Y. Research on patent text classification based on Word2Vec and LSTM. *In*: INTERNATIONAL SYMPOSIUM ON COMPUTATIONAL INTELLIGENCE AND DESIGN, 11., 2018, Hangzhou, China. **Proceedings** […]. IEE, 2018. Sigla do evento: ISCID. p. 71-74. DOI: 10.1109/ISCID.2018.00023.

XU, J.; HE, X.; LI, H. Deep learning for matching in search and recommendation. **Foundations and Trends® in Information Retrieval**, v. 14, n. 2-3, p. 102-288, 2020. Disponível em: http://dx.doi.org/10.1561/1500000076. Acesso em: 9 dez. 2021

YANG, Z. *et al*. XLNet: generlized autoregressive pretraining for language understanding. *In*: CONFERENCE ON NEURAL INFORMATION PROCESSING SYSTEMS, 33., 2019, Vancouver, Canada. **Proceedings** […]. 2019. Sigla do evento: NeurIPS. p. 1-11. Disponível em: https://proceedings.neurips.cc/paper_files/paper/2019/file/dc6a7e655d7e5840e66733e9ee67cc69-Paper.pdf. Acesso em: 3 ago. 2023.

YE, F. *et al*. Cross-domain knowledge discovery based on knowledge graph and patent mining. **Journal of Physics**: Conference Series, v. 1744, 2021. Disponível em: https://iopscience.iop.org/article/10.1088/1742-6596/1744/4/042155/pdf. Acesso em: 7 abr. 2021

YOO, Y. et al. Multi label classification of artificial intelligence related patents using modified D2SBERT and sentence attention mechanism. **arXiv**, 2023. DOI: https://doi.org/10.48550/arXiv.2303.03165.

YU, J. *et al*. A structured representation framework for TRIZ-based Chinese patent classification via reinforcement learning. *In*: INTERNATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE AND BIG DATA, 3., 2020, Chengdu, China. **Proceedings** […]. IEEE, 2020. p. 6-10. Disponível em: https://ieeexplore.ieee.org/document/9137486. Acesso em: 11 ago. 2021

YU, X.; ZHANG, B. Obtaining advantages from technology revolution: a patent roadmap for competition analysis and strategy planning. **Technological Forecasting and Social Change**, v. 145, p. 273-283, 2019.

YU, Z. *et al*. A decision tree method for building energy demand modeling. **Energy and Buildings**, v. 42, n. 10, p. 1637-1646, 2010.

YÜCESOY KAHRAMAN, S.; DERELI, T.; DURMUŞOĞLU, A. Forty years of automated patent classification. **International Journal of Information Technology and Decision Making**, 4 mar. 2023.

YUN, J.; GEUM, Y. Automated classification of patents: a topic modeling approach. **Computers & Industrial Engineering**, v. 147, p. 106636, 2020.

ZHANG, Y. *et al*. Universal adversarial attack via conditional sampling for text classification. **Applied Sciences**, v. 11, n. 20, p. 9539, 2021.

ZHAO, W. X. *et al*. A survey of large language models. **arXiv**, 2023. DOI: https://doi.org/10.48550/arXiv.2303.18223

ZHU, H. *et al*. Patent automatic classification based on symmetric hierarchical convolution neural network. **Symmetry**, v. 12, n. 2, p. 186, 2020.

APÊNDICE A – Seções e classes da taxonomia IPC

Seção	Classe	
A	NECESSIDADES HUMANAS	
	A01	AGRICULTURA; SILVICULTURA; PECUÁRIA; CAÇA; CAPTURA EM ARMADILHAS; PESCA
	A61	CIÊNCIA MÉDICA OU VETERINÁRIA; HIGIENE
	A63	ESPORTES; JOGOS; RECREAÇÃO
B	OPERAÇÕES DE PROCESSAMENTO; TRANSPORTE	
	B01	PROCESSOS OU APARELHOS FÍSICOS OU QUÍMICOS EM GERAL
	B29	PROCESSAMENTO DE MATÉRIAS PLÁSTICAS; PROCESSAMENTO DE SUBSTÂNCIAS EM ESTADO PLÁSTICO EM GERAL
	B32	PRODUTOS EM CAMADAS
	B41	IMPRESSÃO; MÁQUINAS PARA IMPRIMIR LINHAS; MÁQUINAS DE ESCREVER; CARIMBOS
	B65	TRANSPORTE; EMBALAGEM; ARMAZENAMENTO; MANIPULAÇÃO DE MATERIAL DELGADO OU FILAMENTAR
C	QUÍMICA; METALURGIA	
	C07	QUÍMICA ORGÂNICA
	C12	BIOQUÍMICA; CERVEJA; ÁLCOOL; VINHO; VINAGRE; MICROBIOLOGIA; ENZIMOLOGIA; ENGENHARIA GENÉTICA OU DE MUTAÇÃO
E	CONSTRUÇÕES FIXAS	
	E21	PERFURAÇÃO DO SOLO; MINERAÇÃO
F	ENGENHARIA MECÂNICA; ILUMINAÇÃO; AQUECIMENTO; ARMAS; EXPLOSÃO	
	F21	ILUMINAÇÃO
G	FÍSICA	
	G01	MEDIÇÃO; TESTE
	G02	ÓPTICA
	G03	HOROLOGIA
	G06	CÔMPUTO; CÁLCULO OU CONTAGEM
	G08	SINALIZAÇÃO
	G09	EDUCAÇÂO; CRIPTOGRAFIA; APRESENTAÇÃO VISUAL; ANÚNCIOS; LOGOTIPOS
	G11	ARMAZENAMENTO DE INFORMAÇÕES
H	ELECTRICIDADE	
	H01	ELEMENTOS ELÉTRICOS
	H03	CIRCUITOS ELETRÔNICOS
	H04	TÉCNICA DE COMUNICAÇÃO ELÉTRICA
	H05	TÉCNICAS ELÉTRICAS NÃO INCLUÍDAS EM OUTRO LOCAL

APÊNDICE B – Protocolo para revisão integrativa

A estrutura deste protocolo para revisão integrativa da literatura foi adaptada de Kitchenham e Charters (2007) e está organizada da seguinte forma:

1) Data: 10/8/2021
2) Nomes dos pesquisadores/instituições
   a) Luciano Zamperetti Wolski/UFSC
   b) Alexandre Leopoldo Gonçalves/UFSC

3) Fundamentos teóricos da pesquisa:
   a) A fundamentação teórica está descrita no capítulo 2.

4) Questão de pesquisa:
   a) "Quais são os métodos empregados na classificação de documentos de patentes?"

5) Bases de dados consultadas: liste as bases de dados que serão pesquisadas. Liste revistas ou websites que serão pesquisados.
   a) Science Direct®
   b) Scopus®
   c) Web of Science®
   d) IEEExplore®

6) Critérios de inclusão e exclusão:
   a) Critérios de inclusão
      - Os termos de busca devem constar no título, no resumo ou na palavra-chave.
      - Os documentos devem estar preferencialmente em inglês.
      - Os documentos publicados devem respeitar o período de 2017 a 2022.
      - Os documentos devem estar disponíveis para download.

   b) Critérios de exclusão
      - Documentos com inconsistências como falta de título, autor, resumo ou palavras-chave;
      - Documentos duplicados – foram encontrados em mais de uma base.

7) Estratégias de busca:
    a) A expressão utilizada para a busca (“patent classification” OR “patent document classification” OR “patent text classification” OR “patent document categorization”) AND (“machine learning” OR “artificial Intelligence” OR “deep learning” OR “neural network”), onde os termos deveriam aparecer no título, no resumo ou nas palavras-chave do artigo.
    b) A busca por uma expressão com termo composto será utilizada entre aspas (ex.: *“patente classification”*). Também serão utilizados os operadores booleanos (AND, OR) em maiúsculas, caso contrário será considerada parte da expressão de busca. Os parênteses foram utilizados para agrupar termos dentro de uma expressão.

8) Critérios de qualidade para seleção dos artigos
    a) Serão considerados somente os artigos que estão de acordo com a busca realizada e em concordância com a questão de pesquisa.

9) Estratégias de extração dos dados: descrever como os dados serão extraídos dos artigos selecionados.
    a) Será utilizada uma matriz de síntese onde, após a seleção dos artigos, será realizada a leitura dos títulos, dos resumos e das palavras-chave de todas as publicações completas. Em seguida, os artigos selecionados serão relacionados na matriz para posterior análise e adequação aos critérios de inclusão da pesquisa.

10) Estratégias de análise dos dados:
    a) Os dados serão analisados para levantar as lacunas e os métodos utilizados referentes à classificação de patentes.

11) Estratégia de disseminação do conhecimento:
    a) O objetivo é gerar uma seção para a tese e um artigo de revisão integrativa considerando os artigos mais recentes na área sobre classificação de patentes.

12) Cronograma das atividades 2022

Figura B-1 - Cronograma

Mês/Semana / Atividades	Abril				Maio					Junho			
Elaboração do protocolo de pesquisa													
Levantamento dos artigos nas bases predefinidas													
Leitura dos abstracts para confirmação de inclusão do artigo na base de estudo													
Leitura integral dos artigos finais selecionados e anotação na matriz													
Revisão final													
Escrita das seções da tese													

Fonte: elaborado pelo autor (2023)

APÊNDICE C – Detalhamento da versão inicial do modelo

1 MODELO PROPOSTO

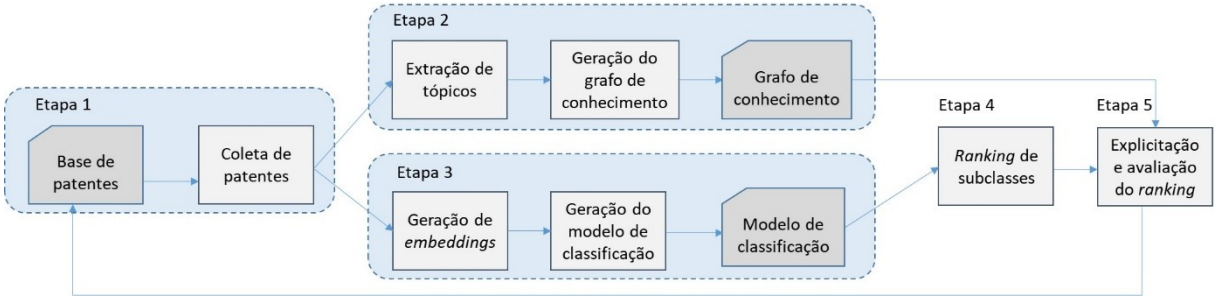
Este apêndice apresenta de forma detalhada o modelo proposto nesta tese com o intuito de facilitar o entendimento, expondo suas etapas e funcionamento. A partir da revisão integrativa da literatura não foram identificados trabalhos que levassem em consideração a sugestão ordenada de subclasses de patentes, tampouco uma forma de explicitar o conhecimento envolto nas sugestões de modo a auxiliar especialistas (examinadores) no processo de análise e identificação das melhores subclasses. Ademais, o modelo também prevê a atualização da base de patentes de modo que o modelo tenha um comportamento dinâmico ao longo do tempo.

Após a apresentação geral do modelo, é realizada uma instanciação, em que são apresentados os componentes tecnológicos visando clarificar todas as etapas e suas interconexões.

1.1 APRESENTAÇÃO DO MODELO

O modelo proposto compreende uma série de etapas com base na revisão integrativa da literatura e na fundamentação teórica, com a finalidade de responder à pergunta de pesquisa e atingir o objetivo geral e os específicos. A seguir, serão apresentadas as etapas do modelo com uma breve descrição, conforme a Figura C-1.

Figura C-1 - Etapas do modelo proposto



Fonte: elaborado pelo autor (2022)

O objetivo final deste trabalho é propor um modelo voltado à sugestão de subclasses de patentes a partir de fontes de dados não estruturados na forma de texto levando em conta aspectos de ordenação de subclasses (*ranking*) e explicitação de conhecimento. Ou seja, o

ranking deve promover subsídios para o examinador de patentes, indicando quais são as subclasses com a maior relevância para serem atribuídas a determinada patente, de forma a auxiliar na tomada de decisão.

A Etapa 1 consiste na escolha de uma base de patentes disponível para realização de testes em formato não estruturado. A coleta de patentes pode ocorrer a partir de diversas fontes, entre elas a base de patentes disponibilizada pelo USPTO®. Após a coleta, torna-se requerida a realização do pré-processamento, por exemplo, a retirada de *stopwords* e a aplicação de lematização.

Na Etapa 2, a partir da base de patentes obtida na Etapa 1, são extraídos os tópicos mais relevantes que servirão de entrada para a geração do grafo de conhecimento, o qual possui como função essencial a explicitação do conhecimento envolto nas sugestões de subclasses a partir de determinada patente de interesse.

Já na Etapa 3, os dados coletados são transformados em vetores densos (*embeddings*) de documentos. Os *embeddings* de documentos são representados na forma de um vetor de características com *n*-dimensões. De modo geral, servem de base para gerar o modelo de classificação que utiliza, por exemplo, uma rede neural para representar os dados de texto, capturando a semântica de palavras e de documentos. Como resultado, tem-se um modelo de classificação na sua essência, mas com um viés de sugestão de subclasses em que cada subclasse possui uma probabilidade associada, criando subsídios para a ordenação (*ranking*) do resultado.

Na Etapa 4, o *ranking* de subclasses de patentes é produzido sob demanda, ou seja, é ativado a partir da demanda de avaliação de alguma nova patente. O *ranking* representa a ordenação das subclasses de patentes geradas no modelo de classificação, ou seja, apresenta as subclasses mais prováveis para uma determinada patente, indo da subclasse mais provável para a menos provável.

Por fim, a Etapa 5 possui como objetivos a explicitação e a inspeção do *ranking* por meio de grafos de conhecimento. No modelo, o KG representa as subclasses de patentes e como estas se interconectam através dos tópicos extraídos a partir das patentes, viabilizando entender de maneira mais adequada as sugestões ordenadas de subclasses. Nesse sentido, este contexto visa prover ferramental relevante aos examinadores de modo a reduzir o tempo de avaliação de uma patente e aumentar a acurácia, ou seja, objetiva fornecer subsídios para decisões mais assertivas. Como ação final, determinada patente retorna para a base agora com as subclasses escolhidas pelo examinador, o que impacta na atualização do modelo de classificação e no grafo de conhecimento, permitindo a evolução das sugestões ao longo do tempo.

1.2 COMPOSIÇÃO DO MODELO

Nas subseções a seguir são apresentados mais detalhes para cada uma das etapas do modelo proposto.

1.2.1 Etapa 1: Coleta de dados e pré-processamento

Como mencionado, esta etapa é responsável pela coleta de patentes. Patentes podem ser coletadas a partir de diferentes fontes de dados abertos na *web,* disponibilizados, por exemplo, pelo USPTO®. Detalhes são providos na seção 3.3.3.1, que apresenta um conjunto de dados de referência em classificação de patentes elaborado por Li *et al.* (2018) com pouco mais de dois milhões de documentos de patentes de utilidade. Os documentos foram reunidos no período de 2006 a 2015, quando foram extraídas as categorias de subclasse, resumo, título e número da patente.

Após a coleta de patentes, os dados passam por uma fase de pré-processamento e transformação, conforme descrito nas seções 3.4.3.2 e 3.4.3.3, respectivamente.

1.2.2 Etapa 2: Geração dos grafos de conhecimento

O grafo de conhecimento é obtido através de técnicas de NLP utilizadas na extração de tópicos e no relacionamento entre estes e as subclasses que representam o domínio de patentes.

Segundo Ye *et al.* (2021), os elementos de conhecimento são divididos em categorias, de acordo com o IPC. Esses elementos serão armazenados em uma base de dados. Após a análise de correlação, é construído o KG geral, representando as subclasses e os seus tópicos associados. Apesar de existir um grafo geral, é possível extrair o KG de cada subclasse ou mesmo de algumas classes, de modo que este sirva de elemento fundamental na explicitação do *ranking* de subclasses de patentes.

1.2.3 Etapa 3: Geração de *embeddings* e do modelo de classificação

Nesta etapa, o conjunto de dados de patentes é transformado em vetores de documentos com a utilização de *embeddings* de documentos a partir dos conteúdos textuais. O primeiro

passo consiste na formação do dicionário de termos distintos obtidos a partir do conjunto de patentes que serão utilizadas na construção do modelo de classificação.

Na sequência, são realizadas operações algébricas, de tal maneira que os vetores de patentes, que possuem alta dimensionalidade, sejam transformados em vetores densos, também chamados de *embeddings.* Os *embeddings* representam as entradas de determinada arquitetura de rede neural que será utilizada no modelo proposto.

Adicionalmente, para cada patente é necessário gerar um vetor representando as subclasses previamente associadas. Nesse sentido, cada patente terá associado um vetor de *n* posições, indicando a quantidade de subclasses do conjunto de dados. Detalhes são apresentados nas seções 3.4.3.2 e 3.4.3.3.

A partir do pré-processamento das patentes e da geração dos vetores, estes podem servir de entrada para determinada arquitetura de rede neural com o intuito de gerar um modelo de classificação e sugestão de subclasses. Assim, qualquer arquitetura que promova suporte a entradas de múltiplas subclasses por patente e que seja capaz de produzir múltiplas saídas indicado a probabilidade de determinada subclasse estar associada a uma patente de interesse pode ser utilizada.

Para a geração do modelo de classificação e sugestão de subclasses, a proposta desta tese se utiliza de uma arquitetura de rede neural profunda com as características já mencionadas e descritas na seção 2.4.1. A primeira fase é constituída por um conjunto de passos, de modo que a rede neural aprenda a representação dos vetores de entrada para as suas subclasses correspondentes. Esse processo é identificado como treinamento e, a cada época (uma época representa a passagem por todo o conjunto de dados), ocorre a validação do que foi aprendido até o presente momento. Uma vez finalizada a fase de treinamento/validação, ocorre a fase de teste, em que documentos de patentes ainda desconhecidos pela rede neural são avaliados. Caso as avaliações sejam adequadas, o modelo neural gerado pode ser utilizado para futuras classificações.

Por fim, utilizando-se do modelo de classificação treinado, determinada patente pode ser avaliada. Considerando uma entrada qualquer composta pela estrutura determinada, o modelo sugere uma lista de prováveis subclasses com seus respectivos pesos, indicando a probabilidade de aquela subclasse estar associada a uma patente de entrada. Esse resultado então serve de subsídio para a Etapa 4.

1.2.4 Etapa 4: Ordenação (*ranking*) de subclasses

Esta etapa é essencial para o modelo visto que facilita a identificação, por parte dos examinadores, da relevância de cada subclasse associada a uma patente. Para tal, conforme mencionado na etapa anterior e considerando uma patente, o modelo de classificação baseado em redes neurais fornece uma lista de subclasses com suas respectivas probabilidades.

Vale mencionar que a soma das probabilidades do vetor de saída em que cada posição se refere a uma subclasse é igual a 1 (um). Nesse sentido, a probabilidade de cada subclasse indica a sua relevância quanto à patente de interesse. Dito isso, percebe-se que o vetor de probabilidades pode ser ordenado de maneira decrescente, ou seja, do maior valor para o menor, objetivando prover uma indicação ordenada, isto é, um *ranking* da representatividade de cada subclasse para uma patente.

1.2.5 Etapa 5: Explicitação e avaliação do *ranking*

Para a avaliação do resultado da etapa anterior, em que se tem como saída a sugestão de um conjunto de subclasses ordenadas (*ranking*) representando a relevância de cada subclasse associada a uma patente, um grafo de conhecimento é utilizado. Ou seja, a partir de cada subclasse sugerida pela etapa anterior o examinador tem acesso ao grafo de conhecimento, que indica em destaque as subclasses mais relevantes sugeridas na etapa anterior, assim como os conceitos que as interconectam. Ademais, o grafo de conhecimento também destaca os tópicos que fazem parte da patente analisada. A possibilidade de investigação dos conceitos envoltos na análise por meio de um grafo de conhecimento objetiva tornar mais claro o resultado produzido pelo *ranking*.

De modo geral, esta etapa tem por finalidade subsidiar a tomada de decisão do examinador na definição final de quais subclasses devem realmente ser atribuídas para determinada patente. Para tal, este possui à sua disposição uma lista de subclasses ordenadas, na forma de um *ranking* e um grafo de conhecimento para explicitar o conhecimento sobre as subclasses sugeridas.

Por fim, o modelo prevê a atualização da base de patentes com as subclasses selecionadas pelo examinador. Ou seja, considerando o conjunto de subclasses definidas como as mais representativas, estas são vinculadas a determinada patente na base de dados. Essa vinculação permite com que o modelo proposto atualize tanto o grafo de conhecimento quanto o modelo de classificação, permitindo um aprendizado constante. Até o momento, essa

característica ainda não foi implementada, tampouco fará parte da avaliação inicial constante neste documento, mas é parte essencial, uma vez que confere ao modelo proposto a capacidade de lidar com aspectos de temporalidade e de dinamicidade.

1.3 INSTANCIAÇÃO DO MODELO

O propósito desta seção é demonstrar, na forma de um exemplo, a instanciação do modelo proposto, visando clarificar como ele se comporta. Dessa forma, pretende-se mostrar os componentes do modelo e a ligação entre eles, assim como as tecnologias envolvidas. O conteúdo textual utilizado para a instanciação das etapas do modelo proposto é apresentado no Quadro C-1, onde a patente US08001811 será considerada como exemplo.

Quadro C1- Exemplificação do conteúdo de uma patente

	Número da patente US08001811
Título	*washing machine having water softening device*
Resumo	*washing machine water softening device improves solubility detergent water softening performance concurrently washing machine includes tub water supply device supplying water tub detergent supply device supplying detergent tub water softening device softening water water softening device disposed water supplied water supply device mixed detergent supplied detergent supply device water mixed detergent supplied water softening device washing machine water softening device*
Subclasse	D06F

Fonte: elaborado pelo autor a partir da patente US08001811 (2022)

1.3.1 Etapa 1: Coleta de dados e pré-processamento

Como visto anteriormente, a base de dados utilizada é a USPTO-2M, que está no formato JSON. O primeiro passo é a transformação dos dados do formato JSON para CSV. O intuito dessa transformação é obter um conjunto menor de dados para aplicar técnicas para extração de características de texto, como remoção de *stopwords*, tokenização e reconhecimento de entidades nomeadas (do inglês: *Named Entity Recognition* - NER).

Para isso, foi utilizada até o momento a linguagem Python®, a qual possui várias bibliotecas que promovem suporte para este estudo. Inicialmente a biblioteca spaCy® foi utilizada, pois possui algumas ferramentas úteis para o NLP. A partir de agora, a referência aos dados será realizada por meio da palavra *corpus*, formado pelas patentes extraídas do conjunto USPTO-2M.

1.3.2 Etapa 2: Geração dos grafos de conhecimento

Após a limpeza dos dados, as entidades da patente US08001811 foram nomeadas utilizando o spaCy® – o resultado pode ser visto no Quadro C-2. O NER identifica as entidades nomeadas, vinculando a essas entidades conceitos relevantes, e realiza o mapeamento para uma base de conhecimento de destino. A base de conhecimento da DBpedia® foi utilizada para efetuar esse mapeamento.

Isso ajuda a lidar com a desambiguação de entidades e com o mapeamento de conceitos para uma base de conhecimento de destino, enriquecendo o grafo de conhecimento por meio de informações sobre os conceitos mapeados da base de conhecimento da DBpedia®.

Quadro C-2 - Representação vetorial utilizando NER

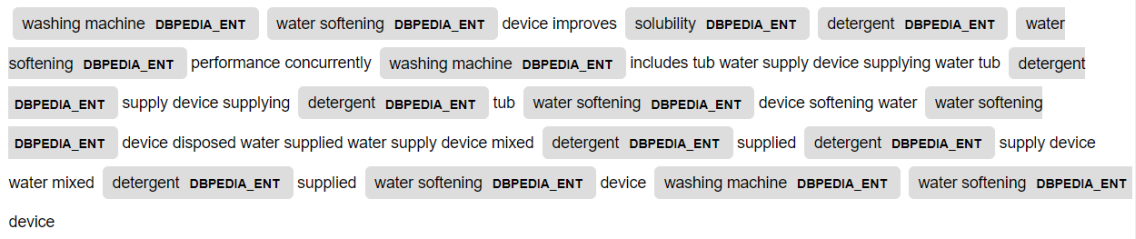
Conceitos		
Descrição	**URI**	**Ocorrências**
detergent	http://dbpedia.org/resource/Detergent	6
solubility	http://dbpedia.org/resource/Solubility	1
washing machine	http://dbpedia.org/resource/Washing_machine	3
water softening	http://dbpedia.org/resource/Water_softening	6

Fonte: elaborado pelo autor (2022)

Foram identificadas 4 entidades *detergent* com 6 ocorrências, *solubility* com 1 ocorrência, *washing machine* com 3 ocorrências e *water softening* com 6 ocorrências. Os conceitos desses termos podem ser acessados pelo URI (*Uniform Resource Identifier*) disponível a partir da base de conhecimento da DBpedia®.

Na Figura C-2, são apresentadas as entidades nomeadas e seus rótulos destacados no texto da patente.

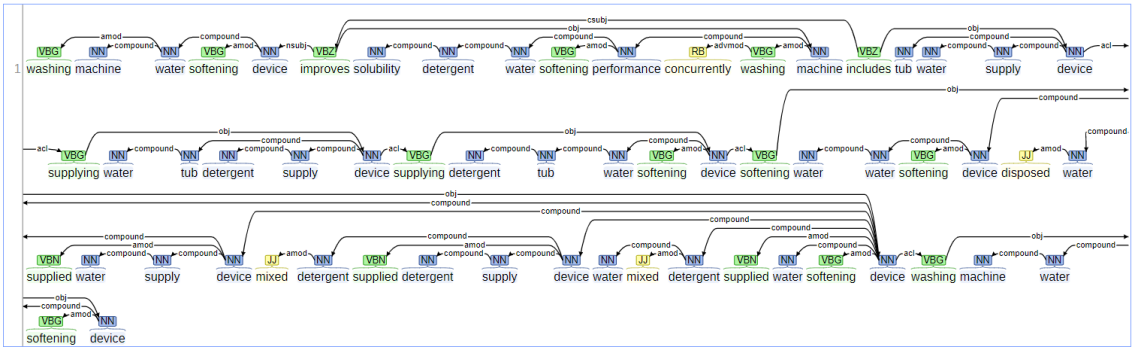
Figura C-2 - Entidades nomeadas e rótulos no texto da patente



Fonte: elaborado pelo autor (2022)

O restante das palavras da lista de *tokens* foi classificado conforme sua subclasse gramatical. As dependências sintáticas entre os *tokens* juntamente com a subclasse gramatical são apresentadas na Figura C-3. Foi utilizada a ferramenta CoreNLP®[28] para uma melhor visualização das dependências. A representação sintática do texto mostra as relações entre os *tokens*. O resultado dessa etapa é um vetor de tópicos que servirá de entrada para a geração do grafo de conhecimento.

Figura C-3 - Dependências sintáticas



Fonte: obtido a partir da patente de exemplo utilizando a ferramenta CoreNLP®

Para a construção do grafo de conhecimento, os elementos mais importantes são os nós e as arestas, onde os nós são representados pelas entidades (tópicos e subclasses) e as arestas pelos relacionamentos. Para gerar o grafo de conhecimento é preciso seguir algumas etapas de forma a converter os dados não estruturados em um grafo. Nesse sentido, cada entidade extraída (considerada aqui como um tópico) é relacionada às subclasses que constam na patente atribuindo a quantidade de vezes (frequência) que os tópicos constam no texto, conforme o Quadro C-3. Nesse sentido, à medida que cada patente é analisada, o grafo evolui. O Quadro C-3 apresenta um exemplo a partir da patente US08001811 com a subclasse D06F e considerando outra patente que possui outra subclasse, aqui identificada como XYZ.

[28] Disponível em: https://stanfordnlp.github.io/CoreNLP/. Acesso em: 28 maio 2023. O CoreNLP® recebe o texto bruto, executa uma série de anotadores em PLN no texto e devolve um conjunto de anotações.

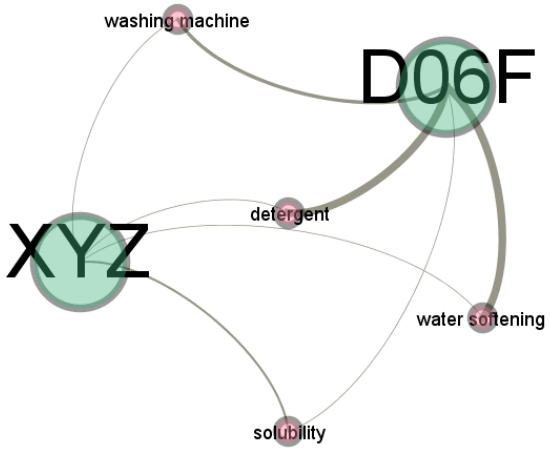
Quadro C-3 - Associação dos tópicos extraídos associados às subclasses

Tópico	Frequências para a subclasse D06F	Frequências para a subclasse XYZ
detergent	6	1
solubility	1	2
washing machine	3	1
water softening	6	1

Fonte: elaborado pelo autor (2022)

O Quadro C-3 indica determinado tópico associado à(s) subclasse(s) com as respectivas frequências. Cada tópico pode estar associado a *n* subclasses. Uma vez que se disponha dos dados integrados, um grafo de conhecimento pode ser gerado conforme a Figura C-4.

Figura C-4 - Grafo de conhecimento com tópicos associados



Fonte: elaborado pelo autor (2022)

Pretende-se com o KG promover a visualização geral dos tópicos do domínio e também os tópicos vinculados a uma ou mais subclasses, como descrito na seção 2.3.1.1. Com o propósito de visualizar esta etapa da instanciação, a ferramenta Gephi® foi utilizada para gerar a representação dos grafos de conhecimento. Essa representação foi criada de forma que seja possível ver as subclasses de patentes, os tópicos dessas subclasses e os relacionamentos entre tópicos e subclasses.

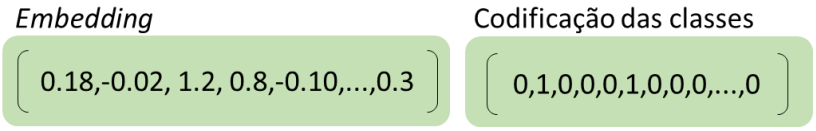
1.3.3 Etapa 3: Geração de *embeddings* e do modelo de classificação

No próximo passo, considera-se a representação de todos os documentos de patentes na forma de um vetor denso. O primeiro passo consiste em transformar os documentos de texto em um vetor de números antes de inserir no modelo. Uma maneira de transformar texto em números é utilizar o *one-hot-encoding* para cada palavra do vocabulário. Após isso, os vetores

codificados de todas as patentes são transformados por meio da aplicação da técnica de *embedding*.

O processo de representação *one-hot-encoding* também é aplicado ao vetor de subclasses que representa cada patente. A Figura C-5 apresenta a representação de entrada de uma patente para a rede neural.

Figura C5 - Entrada de uma patente na rede neural



Fonte: elaborado pelo autor (2022)

O vetor de *embedding* é a transformação de um *vetor one-hot-encoding* em um vetor denso contendo a representação da patente juntamente com o vetor de subclasses. Esses vetores servem como entrada para a geração do modelo de classificação. Por meio das redes neurais MLP, CNN e LSTM, já descritas na seção 2.4.1, o modelo é gerado após as fases de treinamento e validação. Na sequência, na fase de teste, cada patente do vetor de entrada é comparada com as demais patentes que constam na base, permitindo a recuperação (sugestão) de uma lista de subclasses de patentes ordenadas pela relevância.

Para o desenvolvimento do modelo neural, foram utilizadas as bibliotecas TensorFlow® e Keras®, disponíveis na linguagem Python®. Segundo Chollet (2017), a integração entre TensorFlow® e Keras® tornou o aprendizado profundo mais acessível para os usuários.

O TensorFlow® é uma biblioteca de código aberto para aprendizado de máquina que suporta computação numérica baseada em matriz multidimensional, GPU (*Graphics Processing Units*) e processamento distribuído, construção, treinamento e exportação de modelos, entre outros[29] . Já a biblioteca Keras® configura uma estrutura de aprendizado profundo que fornece uma maneira conveniente de definir e de treinar muitas arquiteturas de aprendizado profundo, tendo sido criada com o objetivo de possibilitar uma experimentação rápida (Chollet, 2017).

[29] Disponível em: https://www.tensorflow.org/guide/basics?hl=pt_br. Acesso em: 10 ago. 2021.

1.3.4 Etapa 4: Ordenação (*ranking*) de subclasses

O *ranking* de patentes é representado por uma lista ordenada com as subclasses de patentes mais prováveis geradas pela Etapa 3. O *ranking* de subclasses é produzido por demanda, ou seja, no momento em que surge uma nova patente, ainda sem subclasse. Essa nova patente passa pelas etapas 1, 2 e 3 e, a partir disso, o modelo de classificação oferece quais são as subclasses mais prováveis para a nova patente, possibilitando a geração do *ranking*. O Quadro C-4 apresenta a saída da Etapa 2 e o Quadro C-5 apresenta a saída ordenada. Em ambos os quadros consta a indicação da posição de cada subclasse (considerando as 50 subclasses que compõem o conjunto de dados), a subclasse em si e a probabilidade de cada subclasse.

Quadro C4 - Relação de subclasses sugeridas com as probabilidades

Posição	1	2	...	13	14	15	16	...	49	50
Subclasse	01H	A01N	...	B65D	C07C	C07D	C07H	...	H05B	H05K
Probabilidade	0,0006	0,0294	...	0,0001	0,4900	0,2924	0,0404	...	0,0003	0,0000

Fonte: elaborado pelo autor (2022)

Quadro C-5 - Relação ordenada de classes sugeridas com as probabilidades

Posição	1	2	...	13	14	15	16	...	49	50
Subclasse	C07C	C07D	...	H03M	H01J	B29C	G02B	...	A63F	H01R
Probabilidade	0,4900	0,2924	...	0,0021	0,0019	0,0010	0,0008	...	0,0000	0,0000

Fonte: elaborado pelo autor (2022)

1.3.5 Etapa 5: Explicitação e avaliação do *Ranking*

Uma vez que a lista ordenada de subclasses seja obtida da etapa anterior, o examinador pode inspecionar as subclasses mais relevantes por meio da análise do grafo de conhecimento. Nesse sentido, o examinador pode selecionar uma ou mais subclasses de interesse e solicitar a visualização dessas subclasses por meio do KG, com vistas a auxiliá-lo no entendimento do porquê a subclasse está naquela posição. Mais detalhes sobre o funcionamento dessa etapa constam na seção 5.2.

Após a avaliação das subclasses pelo examinador, ele pode escolher as subclasses mais adequadas à patente que está sendo analisada. Essas subclasses são então atualizadas na base de patentes que suporta o modelo, permitindo a evolução tanto do KG quanto do modelo neural, o que confere ao modelo proposto característica de evolução temporal e dinamicidade. Essa evolução do modelo, bem como a sua avaliação, será prevista na versão final do trabalho.

APÊNDICE D – 50 subclasses mais frequentes

Calculou-se um histograma de subclasses sobre o conjunto de dados total, no qual cada linha continha a subclasse e a quantidade de patentes que mencionavam a subclasse. Foram então selecionadas as primeiras 50 subclasses mais frequentes. Com isso, o conjunto de dados foi inspecionado visando patentes que possuíam determinada subclasse entre as 50 mais frequentes até o limite de 1.000 (mil) patentes, totalizando 50.000 (cinquenta mil). Depois, os dados foram separados em conjuntos de treinamento (40.000) e de teste (10.000), conforme o Quadro D-1.

Quadro D-1 – 50 subclasses mais frequentes

	Subclasse	Descrição
1	A01H	NOVAS PLANTAS OU PROCESSOS PARA OBTÊ-LAS; REPRODUÇÃO DE PLANTAS POR MEIO DE TÉCNICAS DE CULTURA DE TECIDOS
2	A01N	CONSERVAÇÃO DE CORPOS DE SERES HUMANOS OU ANIMAIS OU PLANTAS OU PARTES DESTES; BIOCIDAS, p. ex. COMO DESINFETANTES, COMO PESTICIDAS OU COMO HERBICIDAS; REPELENTES OU ATRATIVOS DE PESTES; REGULADORES DO CRESCIMENTO DE PLANTAS
3	A61B	DIAGNÓSTICO; CIRURGIA; IDENTIFICAÇÃO
4	A61F	FILTROS IMPLANTÁVEIS NOS VASOS SANGUÍNEOS; PRÓTESES; DISPOSITIVOS QUE PROMOVEM DESOBSTRUÇÂO OU PREVINEM COLAPSO DE ESTRUTURAS TUBULARES DO CORPO, p. ex. STENTS; DISPOSITIVOS ORTOPÉDICOS, DE ENFERMAGEM OU ANTICONCEPCIONAIS; FOMENTAÇÃO; TRATAMENTO OU PROTEÇÃO DOS OLHOS OU OUVIDOS; ATADURAS, CURATIVOS OU ALMOFADAS ABSORVENTES; ESTOJOS PARA PRIMEIROS SOCORROS
5	A61K	PREPARAÇÕES PARA FINALIDADES MÉDICAS, ODONTOLÓGICAS OU DE HIGIENE PESSOAL
6	A61M	DISPOSITIVOS PARA INTRODUZIR MATÉRIAS NO CORPO OU DEPOSITÁ-LAS SOBRE ELE; DISPOSITIVOS PARA FAZER CIRCULAR MATÉRIAS NO CORPO OU PARA DELE AS RETIRAR; DISPOSITIVOS PARA PRODUZIR OU POR FIM AO SONO OU À LETARGIA
7	A61N	ELETROTERAPIA; MAGNETOTERAPIA; TERAPIA POR RADIAÇÃO; TERAPIA POR ULTRASSOM
8	A63F	JOGOS DE CARTAS, MESA OU ROLETA; JOGOS EM RECINTOS FECHADOS USANDO PEQUENAS PEÇAS MÓVEIS PARA JOGO; VIDEOGAMES; JOGOS NÃO INCLUÍDOS EM OUTRO LOCAL
9	B01D	SEPARAÇÃO
10	B29C	MOLDAGEM OU UNIÃO DE MATÉRIAS PLÁSTICAS; MOLDAGEM DE MATERIAL EM ESTADO PLÁSTICO, NÃO INCLUÍDO EM OUTRO LOCAL; PÓS-TRATAMENTO DE PRODUTOS MODELADOS, p. ex. REPARO
11	B32B	PRODUTOS EM CAMADAS, i.e. PRODUTOS ESTRUTURADOS COM CAMADAS DE FORMA PLANA OU NÃO PLANA, p. ex. EM FORMA CELULAR OU ALVEOLAR
12	B41J	MÁQUINAS DE ESCREVER; MECANISMOS DE IMPRESSÃO SELETIVA, i.e. MECANISMOS QUE IMPRIMAM DE OUTRA FORMA QUE NÃO A PARTIR DE UMA FORMA; CORREÇÃO DE ERROS TIPOGRÁFICOS

13	B65D	RECIPIENTES PARA ARMAZENAMENTO OU TRANSPORTE DE ARTIGOS OU MATERIAIS, p. ex. SACOS, BARRIS, GARRAFAS, CAIXAS, LATAS, CAIXA DE PAPELÃO, ENGRADADOS, TAMBORES, POTES, TANQUES, ALIMENTADORES, CONTAINERS DE TRANSPORTE; ACESSÓRIOS, FECHAMENTOS OU GUARNIÇÕES PARA OS MESMOS; ELEMENTOS DE EMBALAGEM; PACOTES
14	C07C	COMPOSTOS ACÍCLICOS OU CARBOCÍCLICOS
15	C07D	COMPOSTOS HETEROCÍCLICOS
16	C07H	AÇÚCARES; SEUS DERIVADOS; NUCLEOSÍDEOS; NUCLEOTÍDEOS; ÁCIDOS NUCLEICOS
17	C07K	PEPTÍDEOS
18	C12N	MICRORGANISMOS OU ENZIMAS; SUAS COMPOSIÇÕES; PROPAGAÇÃO, CONSERVAÇÃO OU MANUTENÇÃO DE MICRORGANISMOS; ENGENHARIA GENÉTICA OU DE MUTAÇÕES; MEIOS DE CULTURA
19	C12P	PROCESSOS DE FERMENTAÇÃO OU PROCESSOS QUE UTILIZEM ENZIMAS PARA SINTETIZAR UMA COMPOSIÇÃO OU COMPOSTO QUÍMICO DESEJADO OU PARA SEPARAR ISÔMEROS ÓPTICOS DE UMA MISTURA RACÊMICA
20	C12Q	PROCESSOS DE MEDIÇÃO OU ENSAIO ENVOLVENDO ENZIMAS, ÁCIDOS NUCLEICOS OU MICRO-ORGANISMOS; SUAS COMPOSIÇÕES OU SEUS PAPÉIS DE TESTE; PROCESSOS DE PREPARAÇÃO DESSAS COMPOSIÇÕES; CONTROLE RESPONSIVO A CONDIÇÕES DO MEIO NOS PROCESSOS MICROBIOLÓGICOS OU ENZIMÁTICOS
21	E21B	PERFURAÇÃO DO SOLO OU ROCHA (mineração, exploração de pedreiras E21C; escavação de poços, abertura de galerias ou túneis E21D); OBTENÇÃO DE ÓLEO, GÁS, ÁGUA, MATERIAIS SOLÚVEIS OU FUNDÍVEIS OU UMA LAMA DE MINERAIS DE POÇOS
22	F21V	DETALHES OU CARACTERÍSTICAS DE FUNCIONAMENTO DOS DISPOSITIVOS OU SISTEMAS DE ILUMINAÇÃO; COMBINAÇÕES ESTRUTURAIS DE DISPOSITIVOS DE ILUMINAÇÃO COM OUTROS ARTIGOS, NÃO INCLUÍDOS EM OUTRO LOCAL
23	G01N	INVESTIGAÇÃO OU ANÁLISE DOS MATERIAIS PELA DETERMINAÇÃO DE SUAS PROPRIEDADES QUÍMICAS OU FÍSICAS
24	G01R	MEDIÇÃO DE VARIÁVEIS ELÉTRICAS; MEDIÇÃO DE VARIÁVEIS MAGNÉTICAS
25	G02B	ELEMENTOS, SISTEMAS OU APARELHOS ÓPTICOS
26	G02F	DISPOSITIVOS OU DISPOSIÇÕES ÓPTICOS PARA CONTROLE DE LUZ PELA MODIFICAÇÕES DE PROPRIEDADES ÓPTICAS DO MEIO DOS ELEMENTOS ENVOLVIDOS NOS MESMO; ÓPTICA NÃO LINEAR; MUDANÇA DE FREQUÊNCIA DA LUZ; ELEMENTOS ÓPTICOS LÓGICOS; CONVERSORES ÓPTICOS ANALÓGICOS/DIGITAIS
27	G03B	APARELHOS OU DISPOSIÇÕES PARA TIRAR FOTOGRAFIAS OU PARA PROJETÁ-LAS OU VISUALIZÁ-LAS; APARELHOS OU DISPOSIÇÕES QUE UTILIZAM TÉCNICAS SEMELHANTES POR MEIO DE OUTRAS ONDAS QUE NÃO ONDAS ÓPTICAS; ACESSÓRIOS PARA OS APARELHOS
28	G03G	ELETROGRAFIA; ELETROFOTOGRAFIA; MAGNETOGRAFIA
29	G06F	PROCESSAMENTO ELÉTRICO DE DADOS DIGITAIS (sistemas de computadores baseados em modelos computacionais específicos G06N)
30	G06K	LEITURA DE DADOS GRÁFICOS (reconhecimento ou compreensão de imagem ou vídeo G06V); APRESENTAÇÃO DE DADOS; SUPORTE DE DADOS; MANIPULAÇÃO DE SUPORTE DE DADOS
31	G06Q	TECNOLOGIA DA INFORMAÇÃO E COMUNICAÇÃO [ICT] ESPECIALMENTE ADAPTADA PARA PROPÓSITOS ADMINISTRATIVOS, COMERCIAIS, FINANCEIROS, DE GERENCIAMENTO OU DE SUPERVISÃO; SISTEMAS OU MÉTODOS ESPECIALMENTE ADAPTADOS PARA PROPÓSITOS ADMINISTRATIVOS, COMERCIAIS, FINANCEIROS, DE GERENCIAMENTO OU DE SUPERVISÃO, NÃO INCLUÍDOS EM OUTRO LOCAL
32	G06T	PROCESSAMENTO DE DADOS DE IMAGEM OU GERAÇÃO EM GERAL

33	G08B	SINALIZAÇÃO
34	G09G	DISPOSIÇÕES OU CIRCUITOS PARA CONTROLE DE DISPOSITIVOS INDICADORES UTILIZANDO MEIOS ESTÁTICOS PARA APRESENTAÇÃO DA INFORMAÇÃO VARIÁVEL
35	G11B	ARMAZENAMENTO DE INFORMAÇÕES BASEADO NO MOVIMENTO RELATIVO ENTRE O SUPORTE DE DADOS E O TRANSDUTOR
36	G11C	MEMÓRIAS ESTÁTICAS (dispositivos semicondutores de memória H10B)
37	H01J	VÁLVULAS DE DESCARGA ELÉTRICA OU LÂMPADAS DE DESCARGA
38	H01L	DISPOSITIVOS SEMICONDUTORES NÃO ABRANGIDOS PELA CLASSE
39	H01M	PROCESSOS OU MEIOS, p. ex. BATERIAS, PARA A CONVERSÃO DIRETA DA ENERGIA QUÍMICA EM ENERGIA ELÉTRICA
40	H01R	CONEXÕES ELETROCONDUTORAS; ASSOCIAÇÕES ESTRUTURAIS DE UMA PLURALIDADE DE ELEMENTOS DE CONEXÃO ELÉTRICA MUTUAMENTE ISOLADOS; DISPOSITIVOS DE ACOPLAMENTO; COLETORES DE CORRENTE
41	H03K	TÉCNICAS DIGITAIS
42	H03M	CODIFICAÇÃO, DECODIFICAÇÃO OU CONVERSÃO DE CÓDIGO EM GERAL
43	H04B	TRANSMISSÃO
44	H04J	COMUNICAÇÃO MULTIPLEX
45	H04L	TRANSMISSÃO DE INFORMAÇÃO DIGITAL, p. ex. COMUNICAÇÃO TELEGRÁFICA
46	H04M	COMUNICAÇÃO TELEFÔNICA
47	H04N	COMUNICAÇÃO DE IMAGENS, p. ex. TELEVISÃO
48	H04W	REDES DE COMUNICAÇÃO SEM FIO
49	H05B	AQUECIMENTO ELÉTRICO; FONTES DE LUZ ELÉTRICA NÃO INCLUÍDAS EM OUTRO LOCAL; DISPOSIÇÕES DE CIRCUITOS PARA FONTES DE LUZ ELÉTRICA EM GERAL
50	H05K	CIRCUITOS IMPRESSOS; INVÓLUCROS OU DETALHES ESTRUTURAIS DE APARELHOS ELÉTRICOS; FABRICAÇÃO DE CONJUNTOS DE COMPONENTES ELÉTRICOS

ANEXO A – Patente US08472379

(12) **United States Patent**
Ishii et al.

(10) **Patent No.: US 8,472,379 B2**
(45) **Date of Patent: Jun. 25, 2013**

(54) **MOBILE STATION, RADIO BASE STATION, COMMUNICATION CONTROL METHOD, AND MOBILE COMMUNICATION SYSTEM**

(75) Inventors: **Hiroyuki Ishii**, Yokosuka (JP); **Anil Umesh**, Yokohama (JP)

(73) Assignee: **NTT DoCoMo, Inc.**, Tokyo (JP)

(*) Notice: Subject to any disclaimer, the term of this patent is extended or adjusted under 35 U.S.C. 154(b) by 362 days.

(21) Appl. No.: **12/809,782**

(22) PCT Filed: **Dec. 19, 2008**

(86) PCT No.: **PCT/JP2008/073220**

§ 371 (c)(1),
(2), (4) Date: **Aug. 9, 2010**

(87) PCT Pub. No.: **WO2009/081871**

PCT Pub. Date: **Jul. 2, 2009**

(65) **Prior Publication Data**

US 2010/0296449 A1 Nov. 25, 2010

(30) **Foreign Application Priority Data**

Dec. 20, 2007 (JP) P2007-329125
Dec. 21, 2007 (JP) P2007-331017
Jan. 11, 2008 (JP) P2008-005072

(51) **Int. Cl.**
H04W 4/00 (2009.01)
H04W 36/00 (2009.01)
H04J 3/16 (2006.01)

(52) **U.S. Cl.**
USPC **370/328**; 370/331; 370/465; 455/436

(58) **Field of Classification Search**
None
See application file for complete search history.

(56) **References Cited**

U.S. PATENT DOCUMENTS

7,167,475 B2 * 1/2007 Tourunen et al. 370/394
8,023,460 B2 * 9/2011 Motegi et al. 370/329
2006/0256810 A1 11/2006 Yarlagadda et al.
2009/0116399 A1 * 5/2009 Ho et al. 370/252

FOREIGN PATENT DOCUMENTS

JP 61-075648 A 4/1986
JP 09-214512 A 8/1997
JP 2005-318429 A 11/2005
JP 2007-174120 A 7/2007
JP 2007-180886 A 7/2007
JP 2007-267017 A 10/2007
JP 2007-274658 A 10/2007
WO 01/60017 A1 8/2001
WO 03/105420 A1 12/2003
WO 2007/127558 A2 11/2007

OTHER PUBLICATIONS

Office Action for Japanese Patent Application No. 2009-547082 mailed May 24, 2011, with English translation thereof (6 pages).
3GPP TSG-RAN WG2 #59bis, LG Electronics Inc., R2-074242 "Discussion on RLC Discard", Shanghai, China, Oct. 8, 2007 (3 pages).

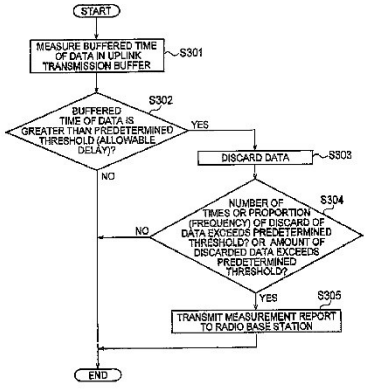
(Continued)

Primary Examiner — Faruk Hamza
Assistant Examiner Cassandra Decker
(74) *Attorney, Agent, or Firm* Osha Liang LLP

(57) **ABSTRACT**

A mobile station according to the present invention includes: a packet discarder unit (**102**, **103**) configured to discard a packet in an uplink transmission buffer, after assigning a sequence number to the packet, when a predetermined condition is met.

5 Claims, 7 Drawing Sheets



ANEXO B – Patente US08394786

(12) **United States Patent**
Freyne et al.

(10) **Patent No.: US 8,394,786 B2**
(45) **Date of Patent: *Mar. 12, 2013**

(54) **QUINAZOLINE DERIVATIVES**

(75) Inventors: **Eddy Jean Edgard Freyne**, Rumst (BE); **Timothy Pietro Suren Perera**, Geel (BE); **Peter Jacobus Johannes Antonius Buijnsters**, Etten-Leur (BE); **Marc Willems**, Vosselaar (BE); **Gaston Stanislas Marcella Diels**, Ravels (BE); **Werner Constant Johan Embrechts**, Beerse (BE); **Peter ten Holte**, Beerse (BE); **Frederik Jan Rita Rombouts**, Antwerpen (BE); **Carsten Schultz-Fademrecht**, Jorvaianica (IL)

(73) Assignee: **Janssen Pharmaceutica N.V.**, Beerse (BE)

(*) Notice: Subject to any disclaimer, the term of this patent is extended or adjusted under 35 U.S.C. 154(b) by 371 days.

This patent is subject to a terminal disclaimer.

(21) Appl. No.: **12/624,637**

(22) Filed: **Nov. 24, 2009**

(65) **Prior Publication Data**

US 2010/0105668 A1 Apr. 29, 2010

Related U.S. Application Data

(62) Division of application No. 10/558,007, filed as application No. PCT/EP2004/005621 on May 25, 2004, now Pat. No. 7,648,975.

(30) **Foreign Application Priority Data**

May 27, 2003 (WO) PCT/EP03/05723
Sep. 15, 2003 (WO) PCT/EP03/10266
Dec. 18, 2003 (WO) PCT/EP03/51061

(51) **Int. Cl.**
A61K 31/33 (2006.01)
C07D 487/00 (2006.01)

(52) **U.S. Cl.** **514/183**; 540/471

(58) **Field of Classification Search** ..... None
See application file for complete search history.

(56) **References Cited**

U.S. PATENT DOCUMENTS

4,067,726	A	1/1978	Sasse et al.
4,160,836	A	7/1979	Vandenberk et al.
6,344,459	B1	2/2002	Bridges et al.
7,648,975	B2	1/2010	Freyne
2002/0173646	A1	11/2002	Thomas et al.

FOREIGN PATENT DOCUMENTS

GB	807899	1/1959
GB	1465451	2/1977
GB	1542514	3/1979
WO	WO 96/09294	3/1996
WO	WO 96/33980	10/1996
WO	WO 96/39145	12/1996

OTHER PUBLICATIONS

U.S. Appl. No. 11/720,693, filed Jun. 1, 2007, Freyne.
Brown, B., et al., "High Throughput Screening—Discovery of Bioactive Substances", Editors: Devlin, John P., Dekker, New York, N.Y, p. 317-328 (1997).
Burke, T., et al. "Protein-Tyrosine Kinase Inhibitors", Drugs of the Future, vol. 17(2) pp. 119-131 (1992).
Davies, S., et al. "Specificity and Mechanism of Action of Some Commonly Used Protein Kinase Inhibitors", Biochem J., vol. 351,pp. 95-106 (2000).
Delia, T., et al. "Fused Pyrimidines, Part Four, Miscellaneous Fused Pyrimidines", Heterocyclic Compounds, John Wiley & Sons, Inc., Interscience Publications, pp. 261-304 (1992).

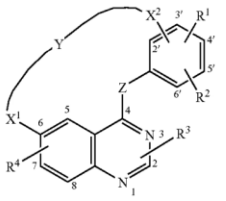
(Continued)

Primary Examiner — Noble Jarrell

(57) **ABSTRACT**

The present invention concerns the compounds of formula

(I)



wherein
Z represents NH;
Y represents —C_{3-9}alkyl-, —C_{2-9}alkenyl-, —C_{3-7}alkyl-CO—NH optionally substituted with amino, mono- or di(C_{1-4}alkyl)amino or C_{1-4}alkyloxycarbonylamino-, —C_{3-7} alkenyl-CO—NH— optionally substituted with amino, mono- or di(C_{1-4}alkyl)amino- or C_{1-4}alkyloxycarbonylamino-, C_{1-5}alkyl-NR^{13}—C_{1-5}alkyl-, —C_{1-5}alkyl-NR^{14}—CO—C_{1-5}alkyl-, —C_{1-6}alkyl-CO—NH—, —C_{1-5}alkyl-$CONR^{15}$—C_{1-5}alkyl-, —C_{1-3}alkyl-NH—CO-Het^{20}-, —C_{1-2}alkyl-CO-Het^{21}-CO—, —C_{1-2}alkyl-NH—CO—$CR^{16}R^{17}$—NH—, —C_{1-2}alkyl-CO—NH—$CR^{18}R^{19}$—CO—, —C_{1-2}alkyl-CO—NR^{20}—C_{1-3}alkyl-CO—, or —NR^{22}—CO—C_{1-3}alkyl-NH—;
X^1 represents a direct bond, O or —O—C_{1-2}alkyl-; X^2 represents a direct bond, —CO—C_{1-2}alkyl-, NR^{12}, —NR^{12}—C_{1-2}alkyl-, —O—N=CH— or —C_{1-2}alkyl-;
R^1 and R^2 are hydrogen or halo;
R^3 are hydrogen; R^4 represents hydrogen or C_{1-4}alkyloxy;
R^{12} and R^{13} are hydrogen or C_{1-4}alkyl;
R^{14} and R^{15} are hydrogen; R^{16} and R^{17} each independently represent hydrogen or C_{1-4}alkyl;
R^{18} and R^{19} are hydrogen or C_{1-4}alkyl optionally substituted with phenyl or hydroxy;
R^{20} and R^{21} are hydrogen or C_{1-4}alkyl optionally substituted with C_{1-4}alkyloxy;
Het^{20}, Het^{21} and Het^{22} are a heterocycle selected from the group consisting pyrrolidinyl, 2-pyrrolidinonyl or piperidinyl optionally substituted with hydroxy.

8 Claims, No Drawings

ANEXO C – Patente USPP22862

US00PP22862P2

(12) **United States Plant Patent**
Eveleens

(10) **Patent No.: US PP22,862 P2**
(45) **Date of Patent: Jul. 17, 2012**

(54) ***GERBERA* PLANT NAMED 'GARORAN'**

(50) Latin Name: ***Gerbera hybrida***
Varietal Denomination: **Garoran**

(75) Inventor: **Jan Leendert Eveleens**, Aalsmeer (NL)

(73) Assignee: **Florist de Kwakel B.V.**, Aalsmeer (NL)

(*) Notice: Subject to any disclaimer, the term of this patent is extended or adjusted under 35 U.S.C. 154(b) by 0 days.

(21) Appl. No.: **12/802,585**

(22) Filed: **Jun. 9, 2010**

(51) **Int. Cl.**
A01H 5/00 (2006.01)

(52) **U.S. Cl.** **Plt./357**

(58) **Field of Classification Search** Plt./357
See application file for complete search history.

(56) **References Cited**

OTHER PUBLICATIONS

Upov-rom GTITM Plant Variety Database 2011/01, GTI Jouve Retrieval Software, Citation for 'Orangina' one page.*

* cited by examiner

Primary Examiner — June Hwu
(74) *Attorney, Agent, or Firm* — C. A. Whealy

(57) **ABSTRACT**

A new and distinct cultivar of *Gerbera* plant named 'Garoran', characterized by its compact, upright and uniformly mounding plant habit; freely flowering habit; orange and yellow bi-colored ray florets; upright and strong scapes; and good garden performance.

1 Drawing Sheet

Botanical designation: *Gerbera hybrida*.
Cultivar denomination: 'GARORAN'.

CROSS-REFERENCED TO CLOSELY-RELATED APPLICATIONS

Title: *Gerbera* Plant Named 'Garsunny'.
Applicant: Jan Leendert Eveleens.
Filed: Jun. 9, 2010.
Ser. No.: 12/802,576.

BACKGROUND OF THE INVENTION

The present invention relates to a new and distinct cultivar of *Gerbera* plant, botanically known as *Gerbera hybrida* and hereinafter referred to by the name 'Garoran'.

The new *Gerbera* plant is a product of a planned breeding program conducted by the Inventor in De Kwakel, The Netherlands. The objective of the breeding program is to create new compact container *Gerbera* plants with numerous inflorescences, good garden performance, frost tolerance and attractive inflorescence coloration.

The new *Gerbera* plant originated from a cross-pollination in March, 2006 in De Kwakel, The Netherlands of a proprietary selection of *Gerbera hybrida* identified as code number 068923, not patented, as the female, or seed, parent with a proprietary selection of *Gerbera hybrida* identified as code number PA 0207, not patented, as the male, or pollen, parent. The new *Gerbera* plant was discovered and selected by the Inventor as a single flowering plant within the progeny of the stated cross-pollination in a controlled greenhouse environment in De Kwakel, The Netherlands during the spring of 2007.

Asexual reproduction of the new *Gerbera* plant by tissue culture in a controlled environment in De Kwakel, The Netherlands since the spring of 2007 has shown that the unique

features of this new *Gerbera* plant are stable and reproduced true to type in successive generations.

SUMMARY OF THE INVENTION

Plants of the new *Gerbera* have not been observed under all possible environmental conditions. The phenotype may vary somewhat with variations in cultural practices and environment such as temperature and light intensity, without, however, any variance in genotype.

The following traits have been repeatedly observed and are determined to be the unique characteristics of 'Garoran'. These characteristics in combination distinguish 'Garoran' as a new and distinct cultivar of *Gerbera* plant:

1. Compact, upright and uniformly mounding plant habit.
2. Freely flowering habit.
3. Orange and yellow bi-colored ray florets.
4. Upright and strong scapes.
5. Good garden performance.

Plants of the new *Gerbera* differ from plants of the female parent selection in the following characteristics:

1. Plants of the new *Gerbera* have smaller inflorescences than plants of the female parent selection.
2. Plants of the new *Gerbera* and the female parent selection differ in ray floret color as plants of the female parent selection have white-colored ray florets.
3. Plants of the new *Gerbera* have shorter scapes than plants of the female parent selection.

Plants of the new *Gerbera* differ from plants of the male parent selection in the following characteristics:

1. Leaves of plants of the new *Gerbera* are longer than leaves of plants of the male parent selection.
2. Plants of the new *Gerbera* and the male parent selection differ in ray floret color as plants of the male parent selection have dark yellow-colored ray florets.

Plants of the new *Gerbera* can be compared to plants of the *Gerbera hybrida* 'Garsunny', disclosed in U.S. Plant patent