ISSN 1983-1838

(DOI): 10.18624/ etech.v18i1.1447

PATENT CLASSIFICATION MODEL BASED ON DENSE VECTOR REPRESENTATION, SORTING TECHNIQUES, AND KNOWLEDGE EXPLICATION

MODELO DE CLASSIFICAÇÃO DE PATENTES BASEADO EM REPRESENTAÇÃO VETORIAL DENSA, TÉCNICAS DE ORDENAÇÃO E EXPLICAÇÃO DO CONHECIMENTO

LUCIANO ZAMPERETTI WOLSKI
https://orcid.org/0000-0003-4683-1013/ lwolski@unemat.br
Universidade do Estado de Mato Grosso – UNEMAT, Mato Grosso, Brasil

ALEXANDRE LEOPOLDO GONÇALVES
https://orcid.org/0000-0002-6583-2807 /alexandre.l.goncalves@gmail.com
Universidade Federal de Santa Catarina – UFSC, Santa Catarina, Brasil

Recebido em: 09/09/2025
Aprovado em: 30/12/2025
Publicado em: 13/02/2026

RESUMO

Este estudo aborda a classificação automática de patentes, visando auxiliar examinadores na categorização eficiente de documentos. O objetivo é propor um modelo que utilize dados não estruturados em texto, levando-se em consideração a ordenação de subclasses e a explicitação de conhecimento. Realizou-se uma revisão integrativa da literatura para identificar métodos apropriados. O modelo é avaliado em dois cenários usando dados do USPTO. No cenário geral, empregando arquiteturas de redes neurais do tipo *transformers*, alcança-se acurácia de aproximadamente 80% na recomendação das 5 primeiras subclasses, considerando 50 documentos recuperados. No cenário específico, comparando-se com redes neurais tradicionais, obtém-se acurácia de 90%. Adicionalmente, explora-se a viabilidade de um grafo de conhecimento para apoiar a explicitação da ordenação de subclasses de patentes, com o intuito de facilitar a explicitação e visualização dos resultados. Os resultados indicam que o modelo proposto pode otimizar o processo de classificação de patentes, facilitando o trabalho dos examinadores na seleção de subclasses adequadas.

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

ABSTRACT

This study addresses the automatic classification of patents, aiming to assist examiners in efficiently categorizing documents. The objective is to propose a model that utilizes unstructured text data, taking into account the ordering of subclasses and the explication of knowledge. An integrative literature review was

ISSN 1983-1838

(DOI): 10.18624/ etech.v18i1.1447

PATENT CLASSIFICATION MODEL BASED ON DENSE VECTOR REPRESENTATION, SORTING TECHNIQUES, AND KNOWLEDGE EXPLICATION

LUCIANO ZAMPERETTI WOLSKI
ALEXANDRE LEOPOLDO GONÇALVES

conducted to identify appropriate methods. The model is evaluated in two scenarios using data from the USPTO. In the general scenario, employing transformer-based neural network architectures, an accuracy of approximately 80% is achieved in recommending the top 5 subclasses, considering 50 retrieved documents. In the specific scenario, compared to traditional neural networks, an accuracy of 90% is obtained. Additionally, the feasibility of a knowledge graph is explored to support the ordering of patent subclasses, with the aim of facilitating the explication and visualization of results. The findings indicate that the proposed model can optimize the patent classification process, making it easier for examiners to select appropriate subclasses.
Keywords: patent analysis; deep learning; patent classification; embedding; knowledge graph

1 INTRODUÇÃO

O cenário global de inovação tem experimentado um crescimento constante, refletido no aumento anual dos pedidos de patentes. Segundo a Organização Mundial da Propriedade Intelectual (WIPO, 2022), em 2021, foram registrados 3,4 milhões de pedidos de patentes em todo o mundo, um acréscimo de 3,6% em relação ao ano anterior. Esses pedidos representam uma vasta coleção de conhecimento humano, geralmente em formato não estruturado, que precisa ser analisada, classificada e gerenciada pelos escritórios de patentes (Risch & Krestel, 2019).

O sistema de classificação mais utilizado internacionalmente é a Classificação Internacional de Patentes (IPC), que divide as tecnologias em oito seções principais e aproximadamente 75 mil subdivisões (WIPO, 2022). Esta estrutura hierárquica visa facilitar a recuperação de documentos similares e auxiliar no processo de busca e análise de patentes. No entanto, o crescente volume de pedidos tem sobrecarregado os examinadores, que tradicionalmente realizam a classificação de forma manual (Sofean, 2021).

Neste contexto, a classificação automática de patentes emerge como um campo de pesquisa crucial. Técnicas de Inteligência Artificial, especialmente o aprendizado de máquina e o aprendizado profundo, têm sido aplicadas para automatizar e otimizar este processo (Yücesoy Kahraman et al., 2023). Contudo, diversos desafios persistem, como a complexidade da linguagem técnica e jurídica das patentes, a constante evolução tecnológica que demanda atualizações frequentes nas classificações, e o imenso volume de dados a ser processado (Yoo et al., 2023).

O problema central abordado neste artigo é a necessidade de desenvolver sistemas de classificação automática de patentes que sejam eficientes, precisos e capazes de lidar com a complexidade e o volume crescente de pedidos de patentes, aliviando assim a carga de trabalho

ISSN 1983-1838

(DOI): 10.18624/ etech.v18i1.1447

dos examinadores e melhorando a qualidade da tomada de decisão nos escritórios de patentes (Jafery et al., 2019).

O objetivo principal desta pesquisa é investigar e propor soluções avançadas para a classificação automática de patentes, explorando técnicas de aprendizado profundo e processamento de linguagem natural. Busca-se desenvolver um modelo que possa classificar patentes com precisão comparável à classificação manual realizada por especialistas, considerando a estrutura hierárquica da IPC e adaptando-se às mudanças tecnológicas ao longo do tempo (Gomez & Moens, 2014).

Diante do contexto apresentado, surge o seguinte problema 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?

Isto posto, apresenta-se neste estudo um modelo voltado a classificação de patentes a partir de texto levando-se em consideração aspectos de ordenação de subclasses e explicitação de conhecimento. Para isso o modelo proposto estabelece uma combinação de técnicas de processamento de linguagem natural (PLN), representação do conhecimento e aprendizagem profunda para auxiliar no processo de classificação de patentes. Mais especificamente, contribuir com o processo de tomada de decisão, de tal maneira que, a tarefa de classificação no contexto de análise de patentes seja facilitada.

2 MODELO PROPOSTO

O modelo proposto consiste em uma sequência de etapas desenvolvidas com base em uma revisão integrativa da literatura e fundamentação teórica. Tem como propósito abordar a questão de pesquisa e alcançar o objetivo estabelecido. As etapas do modelo são apresentadas na Figura 1. Este trabalho tem como meta final a proposição de um modelo para recomendar subclasses de patentes. O modelo utiliza dados não estruturados em formato de texto, extraídos de documentos de patentes. Dois aspectos principais são considerados:

Ordenação (ranking) de subclasses: Visa fornecer aos examinadores de patentes uma lista ordenada de subclasses mais relevantes para uma determinada patente.

ISSN 1983-1838

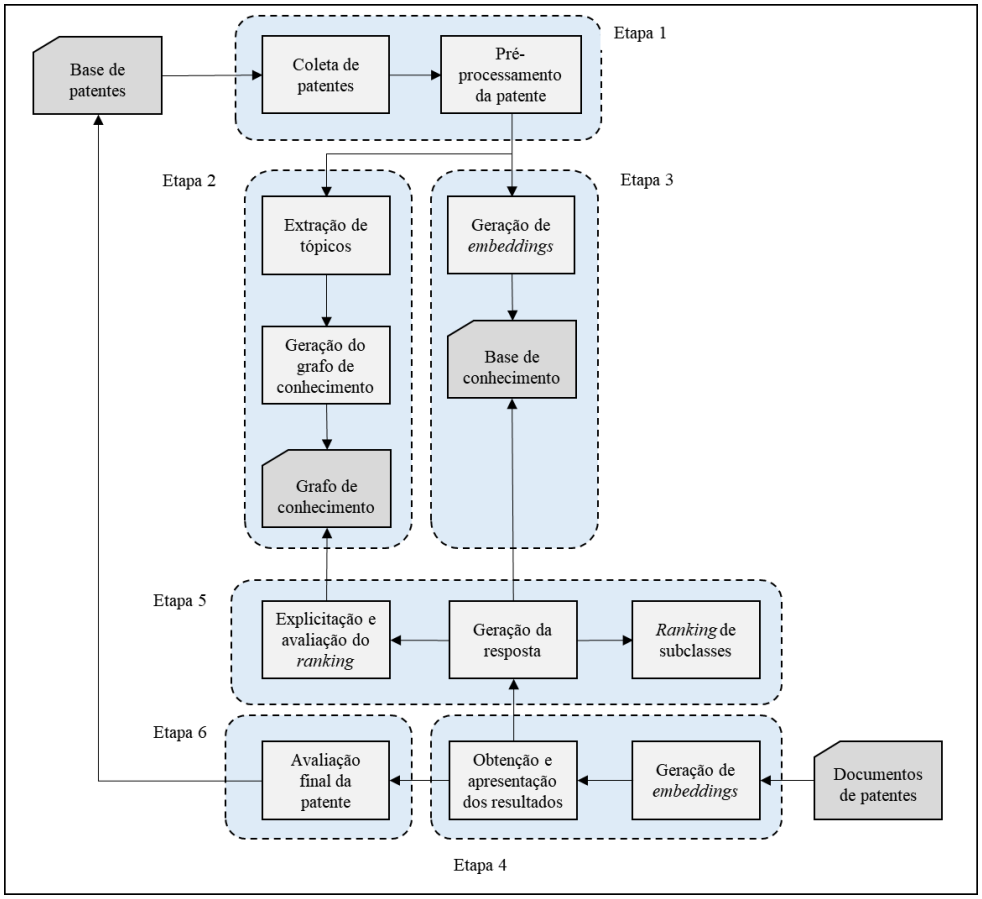
(DOI): 10.18624/ etech.v18i1.1447

Explicitação de conhecimento: Busca evidenciar as relações entre conceitos e as subclasses sugeridas.

O objetivo é auxiliar os examinadores no processo de tomada de decisão, oferecendo tanto uma classificação de relevância quanto uma compreensão clara das conexões conceituais subjacentes.

A Etapa 1 consiste na escolha de uma base de patentes disponível para realização de testes. Os dados de patentes utilizados para a avaliação do modelo foram selecionados da base de patentes americanas USPTO®. Após a coleta, os dados passam por um conjunto de operações de pré-processamento, sendo basicamente utilizada a remoção de eventuais pontuações e *stopwords*.

Figura 1 – Etapas do modelo proposto



Fonte: Do autor (2024)

ISSN 1983-1838

(DOI): 10.18624/ etech.v18i1.1447

Na Etapa 2, são extraídos os tópicos mais relevantes que servirão de entrada para a geração do grafo de conhecimento (KG), 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 ocorre a geração dos *embeddings*, ou seja, os textos das patentes são transformados por diferentes modelos de linguagem pré-treinados (*Pre-Trained Models* - 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.

ISSN 1983-1838

(DOI): 10.18624/ etech.v18i1.1447

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.

3 RESULTADOS E DISCUSSÃO

Esta seção apresenta os resultados obtidos por intermédio da instanciação do modelo proposto, utilizando o conjunto de dados elaborado por Li *et al.* (2018) com pouco mais de dois milhões de documentos de patentes de utilidade, chamado de USPTO-2M. Os documentos foram reunidos no período de 2006 a 2015, sendo compostos por subclasses, resumo, título e número da patente. 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.

ISSN 1983-1838

(DOI): 10.18624/ etech.v18i1.1447

3.1 Ambiente de avaliação do modelo proposto

Visando clarificar o ambiente computacional utilizado neste artigo, utilizou-se uma *workstation* Lenovo P330 equipada com processador Intel Xeon E-2174G, 16 GB de RAM, armazenamento combinado de HD e SSD, e uma Placa Quadro de 2 GB. O estudo avalia quatro modelos pré-treinados (PTMs) baseados em arquiteturas *transformer*, 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 como MLP (*Multilayer Perceptron*), CNN (*Convolutional Neural Network*) e LSTM (*Long Short-Term Memory*).

Os experimentos envolvem testes com *embeddings* de dimensões variadas, aplicando estratégias de *ranking* como Soma das Ocorrências e Soma dos Scores. Os parâmetros *n* (patentes retornadas) e *k* (subclasses sugeridas) são ajustados durante os testes. A acurácia é a métrica principal para avaliação de desempenho.

O conjunto de dados compreende aproximadamente 2 milhões de patentes de 2006 a 2014 para treinamento e cerca de 50 mil patentes de 2015 para teste. Algumas patentes específicas são analisadas individualmente para ilustrar o comportamento do modelo.

O ambiente faz uso de aceleração GPU® através de CUDA® e cuDNN® para otimizar o processamento de redes neurais profundas. O Elasticsearch® é utilizado como banco de dados para armazenamento e indexação das patentes.

Este ambiente abrangente permite uma avaliação detalhada do modelo proposto, comparando diferentes abordagens e utilizando um conjunto de dados substancial de patentes para garantir resultados robustos e representativos.

3.2 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, foram utilizados para geração e indexação dos *embeddings* na base de conhecimento para as patentes dos anos de 2006 a 2014.

ISSN 1983-1838

(DOI): 10.18624/ etech.v18i1.1447

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 1, 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 1 – 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

ISSN 1983-1838

(DOI): 10.18624/ etech.v18i1.1447

LUCIANO ZAMPERETTI WOLSKI
ALEXANDRE LEOPOLDO GONÇALVES

	b) Soma dos *scores*
n	10; 25; 50; 75; 100
k	1; 2; 3; 4; 5; 6; 7; 8; 9; 10

Fonte: Do autor (2024)

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*. 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.

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.

3.2.1 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

ISSN 1983-1838

(DOI): 10.18624/ etech.v18i1.1447

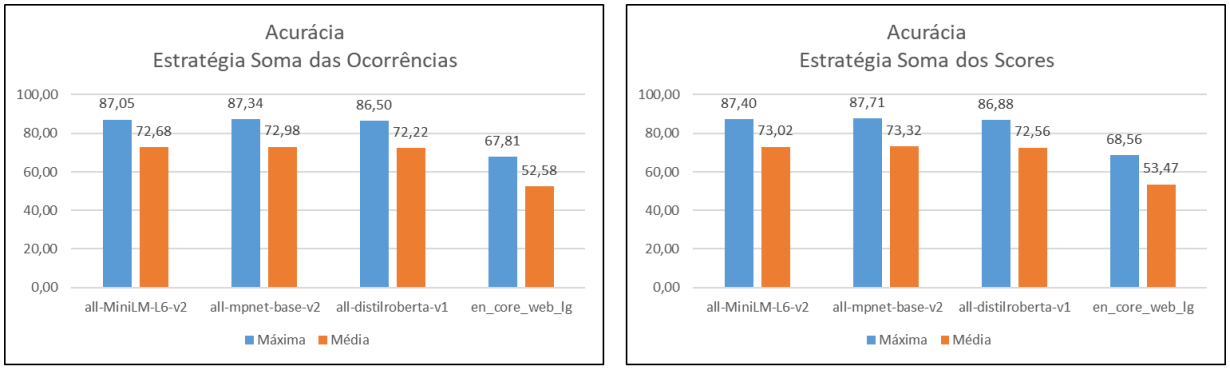
um conjunto de teste objetivando aferir as acurácias nas diferentes instanciações realizadas. Nesta seção, serão detalhados os resultados gerais e específicos levando-se em conta as diferentes combinações de instanciações para gerar as acurácias e, dessa forma, avaliar o modelo.

A acurácia é realizada levando-se em consideração o conjunto de subclasses recomendadas. 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).

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: Do autor (2024)

ISSN 1983-1838

(DOI): 10.18624/ etech.v18i1.1447

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. Os resultados das acurácias dos modelos são levemente superiores com o emprego da estratégia de ordenação SS.

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.

Na Tabela 1 são demonstrados os resultados dos testes realizados para a PTM *all-mpnet-base-v2*, que obteve a melhor acurácia entre os 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).

ISSN 1983-1838

(DOI): 10.18624/ etech.v18i1.1447

LUCIANO ZAMPERETTI WOLSKI
ALEXANDRE LEOPOLDO GONÇALVES

Tabela 1 – 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: Do autor (2024)

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%.

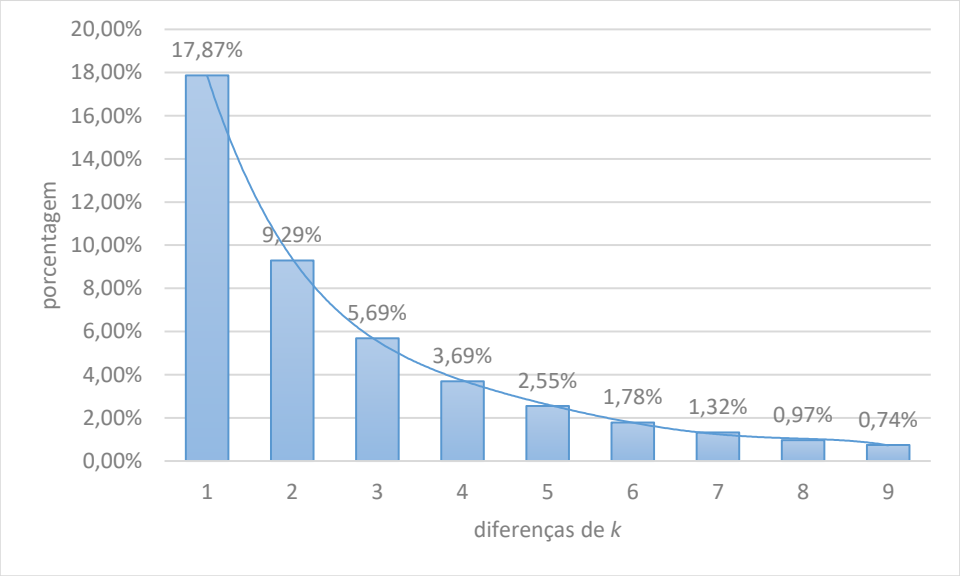
A partir dos resultados, percebe-se que a estratégia de ordenação SS obteve um desempenho um pouco superior em relação à estratégia SO. Quanto maior o valor de *n* e *k*, maior a acurácia do modelo. Sendo assim, à medida que mais subclasses são consideradas, ou seja, que se aumenta a ordem do *ranking*, 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.

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.

ISSN 1983-1838

(DOI): 10.18624/ etech.v18i1.1447

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



Fonte: Do autor (2024)

3.2.2 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 (*Deep Learning)*, 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 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.

ISSN 1983-1838

(DOI): 10.18624/ etech.v18i1.1447

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 deste trabalho, considerando a disponibilidade de dados e os recursos computacionais.

No Quadro 2 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.

ISSN 1983-1838

(DOI): 10.18624/ etech.v18i1.1447

LUCIANO ZAMPERETTI WOLSKI
ALEXANDRE LEOPOLDO GONÇALVES

Quadro 2 - 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: Do autor (2024)

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 3 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 3 – 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: Do autor (2024)

ISSN 1983-1838

(DOI): 10.18624/ etech.v18i1.1447

Para elaboração do Quadro 3, 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 3.

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 2, 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*.

ISSN 1983-1838

(DOI): 10.18624/ etech.v18i1.1447

LUCIANO ZAMPERETTI WOLSKI
ALEXANDRE LEOPOLDO GONÇALVES

Tabela 2 - 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: Do autor (2024)

Nesse sentido, percebe-se pela Tabela 2 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.

ISSN 1983-1838

(DOI): 10.18624/ etech.v18i1.1447

A próxima seção apresenta uma análise individual de uma patente com o objetivo de clarificar o *ranking* utilizando os PTMs e as redes neurais tradicionais. Visto que, 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*.

3.2.2.1 Cenário para a patente nº US08472379

Para clarificar o funcionamento do processo de ordenação das subclasses (*ranking*), o Quadro 4 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.

Quadro 4 - 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: Do autor (2024)

O resultado da execução da etapa de teste da patente de exemplo, para 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%).

ISSN 1983-1838

(DOI): 10.18624/ etech.v18i1.1447

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.

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. A relevância de uma subclasse é calculada dividindo sua frequência nos documentos retornados por 50. Nos PTMs usa-se o conceito de relevância em vez de probabilidade, pois nem todas as subclasses estão presentes nos 50 documentos retornados. Nas redes neurais, a saída é um vetor com probabilidade para todas as 50 subclasses.

Dessa forma, analisando-se 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.

Considerando os resultados, 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%.

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.

ISSN 1983-1838

(DOI): 10.18624/ etech.v18i1.1447

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.

3.3 Grafo de Conhecimento

O presente trabalho utiliza técnicas de NLP (*Natual Language Processing*) na extração de tópicos (nós) e relacionamentos (arestas) entre estes e as subclasses, visando a geração de um grafo de conhecimento (KG - *Knowledge Graph*) 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.

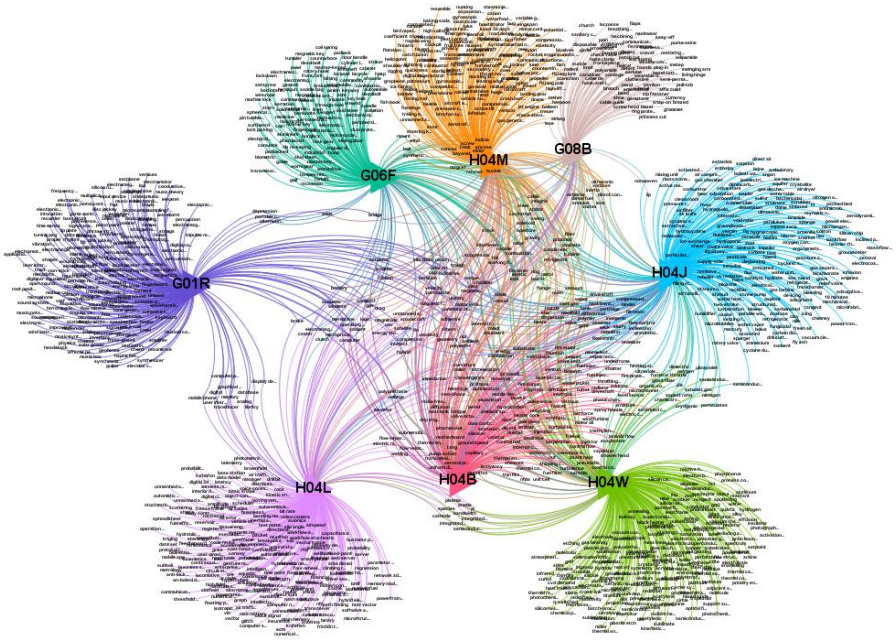
Na2, são apresentadas 8 (oito) subclasses de patentes, indicadas pelos códigos H04W, H04J, H04B, H04L, H04M, G01R, G06F e G08B, com os seus respectivos conceitos.

ISSN 1983-1838

(DOI): 10.18624/ etech.v18i1.1447

LUCIANO ZAMPERETTI WOLSKI
ALEXANDRE LEOPOLDO GONÇALVES

Figura - 2 Grafo de Conhecimento



Fonte: Do autor (2024)

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.

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.

ISSN 1983-1838

(DOI): 10.18624/ etech.v18i1.1447

4 CONCLUSÃO

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.

REFERÊNCIAS

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:

ISSN 1983-1838

(DOI): 10.18624/ etech.v18i1.1447

doi.org/10.1007/978-3-319-12511-4_11.

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.

LI, S. *et al*. DeepPatent: patent classification with convolutional neural networks and word embedding. **Scientometrics**, v. 117, n. 2, p. 721-744, 2018.

RISCH, J.; KRESTEL, R. Domain-specific word embeddings for patent classification. **Data Technologies and Applications**, v. 53, n. 1, p. 108-122, 2019.

SOFEAN, M. Deep learning based pipeline with multichannel inputs for patent classification. **World Patent Information**, v. 66, p. 102060, 2021.

WIPO. **World Intellectual Property Indicators 2022**. Geneva: World Intellectual Property Organization, 2022.

WIPO. **WIPO - Patents**. Disponível em: https://www.wipo.int/patents/en. Acesso em: 11 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.

YÜCESOY KAHRAMAN, S.; DERELI, T.; DURMUŞOĞLU, A. Forty years of automated patent classification. **International Journal of Information Technology and Decision Making**, 4 mar. 2023.