Revista ARACÊ
ISSN: 2358-2472

ANÁLISE DE SIMILARIDADE EM TEXTOS DE PATENTES COM PROCESSAMENTO DE LINGUAGEM NATURAL: REPRESENTAÇÕES VETORIAIS E REDUÇÃO DE DIMENSIONALIDADE - UMA ABORDAGEM VISUAL

SIMILARITY ANALYSIS IN PATENT TEXTS USING NATURAL LANGUAGE PROCESSING: VECTOR REPRESENTATIONS AND DIMENSIONALITY REDUCTION – A VISUAL APPROACH

ANÁLISIS DE SIMILITUD EN TEXTOS DE PATENTES MEDIANTE PROCESAMIENTO DEL LENGUAJE NATURAL: REPRESENTACIONES VECTORIALES Y REDUCCIÓN DE DIMENSIONALIDAD: UN ENFOQUE VISUAL

https://doi.org/10.56238/arev8n6-020

Data de submissão: 05/05/2026 **Data de publicação:** 05/06/2026

Thiago Domingos Marques
Mestre em Propriedade Intelectual e Inovação
Instituição: Universidade Federal de Santa Catarina (UFSC)
E-mail: thiagomestradoufsc@gmail.com

Alexandre Leopoldo Gonçalves
Doutor em Engenharia de Produção
Instituição: Universidade Federal de Santa Catarina (UFSC)
E-mail: a.l.goncalves@ufsc.br

RESUMO
Este artigo apresenta uma abordagem visual inovadora para a análise de relações semânticas entre documentos de patentes, com base em técnicas de Processamento de Linguagem Natural (PLN) e visualização de dados. A análise textual de patentes, especialmente em contextos multilíngues, é fundamental para a pesquisa, a inovação tecnológica e a formulação de políticas públicas. Utilizando TF-IDF (Term Frequency-Inverse Document Frequency), similaridade do cosseno e o algoritmo de agrupamento K-Means, no corpo textual de 720 documentos de patentes, bem como, utilizando vetorização e técnicas de redução de dimensionalidade PCA (Principal Component Analysis) e t-SNE (t-Distributed Stochastic Neighbor Embedding), estruturamos semanticamente um conjunto de patentes depositadas em diferentes idiomas. A modelagem de redes e sua representação gráfica permitiram identificar agrupamentos temáticos e áreas tecnológicas críticas, evidenciando padrões ocultos nos dados. Os resultados demonstram o potencial das visualizações interativas como ferramentas estratégicas para a gestão da informação tecnológica, auxiliando examinadores, pesquisadores e formuladores de políticas. Além disso, o estudo revela desafios específicos enfrentados por determinadas tecnologias no processo de concessão de patentes, contribuindo para uma compreensão mais aprofundada do ecossistema de inovação e propriedade intelectual. Este estudo contribui para os estudos de linguagem técnica e científica ao aplicar representações vetoriais para mapear semelhanças semânticas entre resumos de patentes, revelando padrões linguísticos específicos de áreas tecnológicas.

Palavras-chave: Patentes. Embeddings Textuais. PCA e t-SNE. Processamento de Linguagem Natural. Inovação Tecnológica.

Revista ARACÊ
ISSN: 2358-2472

ABSTRACT
This article presents an innovative visual approach to analyzing semantic relationships among patent documents, based on Natural Language Processing (NLP) techniques and data visualization. Textual analysis of patents, especially in multilingual contexts, is fundamental for research, technological innovation, and public policy development. Using TF-IDF (Term Frequency–Inverse Document Frequency), cosine similarity, and the K-Means clustering algorithm on the textual body of 720 patent documents, as well as applying vectorization and dimensionality reduction techniques such as PCA (Principal Component Analysis) and t-SNE (t-Distributed Stochastic Neighbor Embedding), we semantically structured a set of patents filed in different languages. Network modeling and its graphical representation enabled the identification of thematic clusters and critical technological areas, highlighting hidden patterns within the data. The results demonstrate the potential of interactive visualizations as strategic tools for managing technological information, supporting examiners, researchers, and policymakers. Furthermore, the study reveals specific challenges faced by certain technologies in the patent granting process, contributing to a deeper understanding of the innovation and intellectual property ecosystem. This study also contributes to research on technical and scientific language by applying vector representations to map semantic similarities among patent abstracts, uncovering linguistic patterns specific to technological domains**.**

Keywords: Patents. Textual Embeddings. PCA and t-SNE. Natural Language Processing. Technological Innovation.

RESUMEN
Este artículo presenta un enfoque visual innovador para el análisis de las relaciones semánticas entre documentos de patentes, basado en técnicas de procesamiento del lenguaje natural (PLN) y visualización de datos. El análisis textual de patentes, especialmente en contextos multilingües, es fundamental para la investigación, la innovación tecnológica y la formulación de políticas públicas. Mediante TF-IDF (frecuencia de término-frecuencia inversa de documento), similitud de coseno y el algoritmo de agrupamiento K-Means aplicado al texto de 720 documentos de patentes, así como mediante técnicas de vectorización y reducción de dimensionalidad como PCA (análisis de componentes principales) y t-SNE (incrustación estocástica de vecinos distribuidos en t), se estructuró semánticamente un conjunto de patentes presentadas en diferentes idiomas. El modelado de redes y su representación gráfica permitieron identificar grupos temáticos y áreas tecnológicas críticas, destacando patrones ocultos en los datos. Los resultados demuestran el potencial de las visualizaciones interactivas como herramientas estratégicas para la gestión de la información tecnológica, facilitando el trabajo de examinadores, investigadores y responsables políticos. Además, el estudio revela desafíos específicos que enfrentan ciertas tecnologías en el proceso de concesión de patentes, contribuyendo a una comprensión más profunda del ecosistema de innovación y propiedad intelectual. Este estudio contribuye al análisis del lenguaje técnico y científico mediante la aplicación de representaciones vectoriales para mapear similitudes semánticas entre resúmenes de patentes, revelando patrones lingüísticos específicos en áreas tecnológicas.

Palabras clave: Patentes. Incrustaciones Textuales. PCA y t-SNE. Procesamiento del Lenguaje Natural. Innovación Tecnológica.

Revista ARACÊ
ISSN: 2358-2472

1 INTRODUÇÃO

Este artigo apresenta uma análise com uso de dados de patentes, para análise de similaridades entre esses documentos. O aumento exponencial do número de patentes registradas globalmente, revela a necessidade fundamental de dispor de metodologias eficientes para organizar, comparar e analisar grandes volumes de dados textuais (Krestel *et al.,* 2021; Wolski *et al.,* 2022). Neste contexto, técnicas de mineração de texto e aprendizado de máquina desempenham um papel crucial na identificação de padrões de similaridade entre documentos.

O sistema de patentes exerce um papel central na proteção de invenções e no estímulo à inovação. Cada pedido de patente é submetido a um processo administrativo que culmina em decisões como concessão e rejeição (Marques; Gonçalves, 2025). Neste estudo, a construção de diferentes gráficos é utilizada para análise de similaridades entre patentes, com objetivo de subsidiar a análise de anterioridade (Necessárias na verificação de patentes, para confirmar seu ineditismo e ausência de depósitos anteriores do mesmo produto).

Além de garantir proteção jurídica às invenções, a análise de patentes é estratégica para o desenvolvimento tecnológico, pois permite identificar tendências e retornos sobre investimentos em pesquisa e inovação. No entanto, o volume crescente de pedidos, como destacam Krestel *et al.* (2021), Wolski *et al.* (2022) e Zhang, Guo e Lu (2024), impõe desafios significativos à sua avaliação, exigindo abordagens computacionais mais robustas.

A crescente disponibilidade de dados em múltiplos idiomas apresenta desafios significativos para tarefas de análise e classificação. Métodos tradicionais de processamento textual, como a Representação *Bag-of-Words*, frequentemente falham em capturar o significado semântico e as relações contextuais entre palavras, especialmente em contextos multilíngues.

A visualização em rede desses dados permite identificar áreas tecnológicas com maior ou menor propensão à concessão de patentes, constituindo recurso valioso para examinadores, pesquisadores e formuladores de políticas. Este estudo propõe-se a responder às seguintes questões centrais: Como técnicas de PLN e visualização de dados podem revelar padrões de similaridade semântica entre textos de patentes multilíngues?

O presente trabalho justifica-se pela relevância técnica e econômica de identificar tendências e padrões na distribuição e nas similaridades observadas nos resumos de patentes, bem como de analisar suas relações com os diferentes estágios do processo de patenteamento.

Nesse sentido, este estudo contribui para o avanço das investigações sobre linguagem técnica e científica ao empregar representações vetoriais para mapear semelhanças semânticas entre resumos

Revista ARACÊ
ISSN: 2358-2472

de patentes, e, por conseguinte, evidenciando padrões linguísticos característicos de diferentes áreas tecnológicas.

O objetivo principal é analisar padrões de similaridade semântica entre documentos de patentes por meio de técnicas de vetorização textual e visualização de dados. Para tanto, propõe-se como objetivos específicos são: a) Identificar padrões de similaridade semântica entre os resumos de patentes em diferentes idiomas; b) Aplicar técnicas de vetorização textual (TF-IDF e *Embeddings*) e redução de dimensionalidade (PCA e t-SNE) para agrupar os documentos; c) Desenvolver visualizações interativas que evidenciem relações temáticas entre os documentos analisados.

2 REFERENCIAL TEÓRICO

O estudo das patentes como fontes estratégicas de informação tecnológica tem se consolidado como ferramenta essencial para compreender a dinâmica da inovação e orientar políticas públicas e estratégias empresariais. A análise de documentos de patentes — potencializada por redes semânticas, teoria dos grafos e ferramentas computacionais — permite identificar tendências emergentes, padrões colaborativos e fluxos de conhecimento em diversos setores. O uso de visualizações gráficas e bancos relacionais, amplia o entendimento sobre a evolução tecnológica e o desempenho inovador de empresas e instituições (Krestel *et al.,* 2021; Wolski *et al.,* 2022; Zhang, Guo e Lu, 2024).

Entretanto, a análise automatizada enfrenta desafios específicos, Risch e Krestel (2019) destacam que a linguagem técnica e especializada utilizada nos documentos de patente representa um grande desafio para a automação da classificação, tarefa essencial para a análise de novidades e comparação entre pedidos. Para enfrentar essa dificuldade, os autores propuseram *embeddings* de palavras específicas do domínio, treinados com mais de cinco milhões de patentes, e integrados a um modelo de aprendizado profundo com unidades recorrentes (GRU).

As patentes configuram-se como fontes públicas, confiáveis e estratégicas de informação para a prospecção de tendências tecnológicas e o entendimento da dinâmica da inovação. De acordo com Alves, Souza e Neder (2022), esses registros funcionam como repositórios valiosos, cuja análise, por meio de redes semânticas, permite identificar tendências emergentes e imergentes — como aprendizado de máquina e coleta de dados — e visualizar, de forma estruturada, a evolução tecnológica e suas aplicações.

Krestel *et al.* (2021) corroboram essa perspectiva ao apontarem que as coleções de documentos de patentes representam uma fonte inestimável de conhecimento, embora seu crescimento exponencial imponha desafios à recuperação e análise eficiente desses dados. Nascimento *et al.* (2019) complementam que, na sociedade do conhecimento, o maior desafio já não

Revista ARACÊ
ISSN: 2358-2472

é o acesso à informação, mas a seleção e interpretação crítica, de modo a transformá-la em conhecimento estratégico. Nesse sentido, a busca eficaz em bancos de patentes exige, como defendem Villa e Wirz (2022), a identificação das classes classificatórias mais relevantes, seguida da avaliação visual dos documentos. Contudo, o aumento do volume e da interdisciplinaridade dos pedidos de patentes torna esse processo cada vez mais complexo, elevando o esforço técnico necessário para análise (Krestel *et al.,* 2021; Wolski *et al.,* 2022; Marques; Gonçalves, 2025).

Diante desse cenário, torna-se imprescindível conhecer o estado da arte em nível global antes da formulação de novos projetos de pesquisa, a fim de evitar redundâncias e promover inovações verdadeiramente originais (Brito; Ozaki, 2019). Para Wolski et al. (2022), o sistema de patentes se constitui como um instrumento essencial para o avanço da inovação e o crescimento econômico, ainda que enfrente desafios decorrentes da natureza cumulativa e interconectada das tecnologias digitais. A era digital, portanto, impõe tanto novas possibilidades quanto obstáculos ao sistema de patentes, exigindo abordagens analíticas mais robustas (Wolski *el at.,* 2022; Marques; Gonçalves 2023).

Segundo Ouyang et al. (2022), o sistema de patentes constitui um instrumento essencial para impulsionar a inovação e o crescimento econômico das nações. Contudo, a natureza cumulativa e interconectada das tecnologias digitais impõe desafios adicionais à sua efetividade. Nesse contexto, a era digital não apenas amplia as possibilidades de aplicação do sistema de patentes, como também introduz novas oportunidades e, simultaneamente, novos desafios.

Nesse contexto, a análise estratégica das redes de patentes emerge como ferramenta indispensável. A construção dessas redes com base em relações como coautorias, co-titularidades e citações permite revelar padrões colaborativos, rotas tecnológicas e atores-chave no ecossistema da inovação. Bhatt et al. (2023) argumentam que essas estruturas evidenciam não apenas os vínculos entre agentes inovadores, mas também a direção e intensidade dos fluxos de conhecimento, sendo especialmente úteis em setores disruptivos.

Do mesmo modo, Zhang, Guo e Lu (2024) demonstram que o mapeamento das redes de colaboração permite visualizar a concentração do conhecimento, a centralidade de determinados atores e identificar lacunas na cooperação. Miao *et al.* (2024) propõem um modelo que integra aspectos semânticos e topológicos para analisar a formação dinâmica de redes de citações entre patentes, permitindo identificar com maior precisão os padrões evolutivos da inovação tecnológica.

Sob essa ótica, a participação em redes de colaboração internacional torna-se estratégica. Moresi, Pinho e Hedler (2022) destacam que as patentes refletem áreas de maior interesse dos inventores, sendo cruciais para compreender movimentos de inovação em distintos campos.

Revista ARACÊ
ISSN: 2358-2472

Essa relevância se amplifica no contexto de redes internacionais, nas quais a articulação entre empresas, universidades e centros de pesquisa possibilita o compartilhamento técnico, o acesso a mercados e a atração de investimentos.

Complementariamente, Alonso-Martínez, González-Álvarez e Nieto (2021) apontam que a inserção em redes transnacionais está associada ao fortalecimento de capacidades tecnológicas, ao aumento das atividades empreendedoras, à diversificação de produtos e à transferência de tecnologia — fatores centrais para o desempenho inovador de startups e empresas consolidadas.

De acordo com Liu *et al.* (2024), patentes submetidas mais recentemente apresentam uma velocidade maior de transformação, o que indica um dinamismo crescente na inovação tecnológica contemporânea. Os autores também destacam que as características dinâmicas estruturadas exercem influência mais significativa na previsão de transformação de patentes do que os atributos estáticos, evidenciando a importância de variáveis temporais (Liu *et al.,* 2024).

Conforme Pires et al. (2020), os sistemas de busca de patentes — inclusive aqueles de acesso gratuito — têm evoluído continuamente, incorporando novas ferramentas para a busca e a análise de informações tecnológicas. No campo da verificação e análise de similaridade, a inteligência artificial pode ser empregada para identificar patentes com conteúdo técnico correlato, servindo como base para pesquisas futuras. Tal análise pode utilizar métodos como aprendizado de máquina não supervisionado ou processamento de linguagem natural, nos quais as patentes são representadas por vetores numéricos e a similaridade é aferida a partir da proximidade entre esses vetores (Pires, 2020).

Nesse sentido, a utilização de ferramentas computacionais tem potencializado a análise de patentes. Damo (2021) observa que tais recursos permitem a extração automatizada de dados e sua organização em bancos relacionais, possibilitando a criação de gráficos, dashboards e mapas de redes tecnológicas. Essa abordagem facilita o mapeamento do desenvolvimento tecnológico, revelando padrões de inovação, citações entre patentes e transbordamentos de conhecimento entre áreas distintas.

A visualização gráfica dessas conexões também possui um papel comunicacional relevante. Paulino (2020) interpreta os grafos como imagens complexas, dotadas de valor informacional e simbólico. A partir de estudos sobre mobilizações públicas na plataforma X (Twitter) durante a pandemia de Covid-19, evidencia-se que os grafos são capazes de revelar conexões e discursos que não emergem nas mídias tradicionais (Paulino, 2020). Assim, tais representações podem ser transpostas para o contexto da análise de documentos de patentes, contribuindo para a visualização de relações e fluxos de conhecimento ocultos.

Por meio da aplicação da teoria dos grafos e da ciência das redes complexas, torna-se possível

Revista ARACÊ
ISSN: 2358-2472

estabelecer correlações entre diferentes documentos de patentes, aprimorando o entendimento e o fluxo do processo patentário, com ganhos de desempenho e eficiência. Tal abordagem fornece subsídios relevantes para a formulação de políticas públicas mais eficazes na área de propriedade intelectual (INPI, 2025; Marques; Gonçalves, 2025). Compreender a estrutura e a dinâmica dessas redes, portanto, torna-se essencial tanto para a definição de estratégias empresariais competitivas quanto para o desenvolvimento de iniciativas institucionais que promovam a inovação sistêmica (INPI, 2025).

A verificação em rede é essencial para uma melhor análise de dados e informações. Nessa seara, Jiang *et al.,* (2023) desenvolvem um modelo de aprendizado profundo que combina informações textuais e estruturais de redes para prever o resultado de pedidos de patentes, demonstrando que a integração de embeddings de texto e rede melhora significativamente a precisão das previsões em relação a métodos convencionais.

Nesse contexto, para melhor visualização de informações, segundo Liu *et al.* (2024), o uso de gráficos dinâmicos para representar características temporais de patentes melhora substancialmente a acurácia na previsão de transformação tecnológica ao longo dos anos. O estudo realizado pelos autores aponta que mais de 90% das transformações ocorrem dentro de um período de treze anos após o registro da patente, demonstrando padrões temporais significativos no processo de inovação (Liu *et al.,* 2024).

Segundo Witschard *et al.* (2022), a combinação de múltiplos *embeddings* pode melhorar significativamente a análise de similaridade textual. Os autores sugerem que o uso de ferramentas de visualização analítica interativa pode ajudar analistas humanos a compreender e otimizar modelos de similaridade textual, promovendo maior interpretabilidade e confiança nos resultados.

3 METODOLOGIA

Quanto à dimensão e à classificação, trata-se de uma pesquisa aplicada, de natureza documental e com abordagem quali-quantitativa, tendo como finalidade a descrição do fenômeno estudado (Quadro 01).

Os procedimentos técnicos envolvem pesquisa documental e o uso de técnicas computacionais, como Processamento de Linguagem Natural (PLN), análise de redes e modelagem preditiva. A abordagem metodológica adotada é mista (Quadro 01).

Revista ARACÊ
ISSN: 2358-2472

Quadro 1: Metodologia.

Dimensão	Classificação
Tipo de pesquisa	Aplicada; documental; quali-quantitativa
Finalidade	Descritiva
Procedimentos técnicos	Pesquisa documental; técnicas computacionais (NLP, análise de redes, modelagem preditiva)
Abordagem	Mista

Fonte: Autores (2025).

A escolha das técnicas de Processamento de Linguagem Natural (PLN) e de aprendizado de máquina, nesse trabalho, em especial o uso de representações vetoriais (embeddings), justifica-se pela capacidade dessas abordagens de capturar relações semânticas sutis entre termos e expressões presentes nos resumos dos diferentes documentos de patentes. Sendo que, essa característica é fundamental para atingir o objetivo do estudo, que consiste em identificar padrões linguísticos e temáticos, bem como mapear semelhanças entre documentos, possibilitando uma análise mais precisa e estratégica do panorama tecnológico e de suas tendências e relações.

3.1 CONJUNTO DE DADOS

Foi utilizado como objeto de coleta de amostragem para produção do protótipo, 720 resumos de patentes (Figura 1) retiradas de um Dataset do banco de dados do site: https://www.kaggle.com/

Figura 1: Dataset de Resumos de Patentes.

	abs_Chinese	abs_English	abs_Original
0	本发明提供了一种记录装置(100)和一种便携式电子装置(1000)。记录装置(100)包括...	The invention provides a recording device (100...	Sáng chế đề xuất bộ máy ghi hình (100) và thiế...
1	本发明提供了一种记录装置(200)和一种便携式电子装置(300)。记录装置(200)包括至...	The invention provides a recording device (200...	Sáng chế đề xuất bộ máy ghi hình (200) và thiế...
2	本发明提供了一种记录装置(100)和一种便携式电子装置(1000)。记录装置(100)包括...	The invention provides a recording device (100...	Sáng chế đề xuất bộ máy ghi hình (100) và thiế...
3	本发明涉及一种基于在玻璃基板(1)平面上生成金属Ag纳米颗粒修饰的SnO2纳米线网络自热效应...	The invention relates to an H2S gas sensor bas...	Sáng chế đề cập đến cảm biến khí H2S dựa trên ...
4	本发明涉及通过载波选择至少一个资源以便在子链路发送至少一个分组的用户设备和方法。用户设备包...	The invention relates to a user equipment and ...	Sáng chế đề cập đến thiết bị người dùng và phư...
...	...	...	...
715	修订23/01/2017，提供的是促进多巴胺、去甲肾上腺素等脑内单胺释放并提供预防脑神经疾病...	Revision 23/01/2017 provides a food additive m...	แก้ไข 23/01/2560 ที่จัดให้มีคือวัตถุเติมสำหรับส...
716	本发明涉及在肉类或水产加工食品的生产过程中，利用转谷氨酰胺酶和具有糖类代谢活性的酶将A-1，...	The invention relates to a method for producin...	Sáng chế đề cập tới phương pháp sản xuất thực ...
717	本发明是一种满足由稀释米粉制成直径，厚度合理的圆形包衣饼生产技术的圆形包衣饼生产设备。通过...	The invention relates to a round coated cake p...	Sáng chế đề cập đến thiết bị sản xuất bánh trá...
718	新请求更新于2016年2月26日，含有乳制品成分的食品成分。并含有与蜡质卡扎因-希比泰淀粉酶...	The new request was updated on Feb. 26, 2016 ...	คำขอใหม่ปรับปรุง วันที่ 26/02/2559 ของ...
719	一种具有多个小孔的食品容器，用于将液体，特别是脂肪从食品中分离出来。其特征是在容器(1)的底...	The utility model relates to a food container ...	ภาชนะใส่อาหารที่มีหลุมขนาดเล็กจำนวนหนึ่งสำหรับ...

720 rows × 4 columns

Fonte: Autores (2025).

Utilizou-se um dataset multilíngue de domínio público, extraído do arquivo *Multilingual_classification_dataset.xlsx*, contendo metadados e dados de 720 pedidos de patentes

Revista ARACÊ
ISSN: 2358-2472

registrados em múltiplas jurisdições. proveniente de patentes resgatadas via base de dados de patentes da *Kaggle*, sendo importado em formato de arquivo ".*csv*".

Foi feita aquisição e tratamento de dados, onde os dados foram organizados em arquivos, com campos como resumo. O pré-processamento textual incluiu remoção de *stopwords*, pontuação e normalização de caracteres, seguido por uma filtragem para manter exclusivamente os resumos redigidos em inglês, a fim de garantir a homogeneidade linguística da amostra.

A parte inicial do código, utilizando o mesmo banco de dados, foi objeto de artigo científico publicado com o título "Descobrindo Conexões e Similaridades em Textos de Patentes: Processamento de Linguagem Natural e Visualização Interativa", com enfoque na análise de redes e nas conexões estabelecidas entre os documentos de patentes.

A escolha pela análise dos resumos justifica-se pela agilidade na identificação preliminar das categorias tecnológicas. Essa abordagem permite identificar quais áreas tecnológicas receberam maior volume de pedidos, para identificação das similaridades nos documentos.

3.2 PROCESSAMENTO DOS DADOS

O projeto foi desenvolvido na plataforma *Google Colaboratory*, aproveitando sua integração com o *Google Drive* para armazenamento e acesso aos dados. Diversas bibliotecas da linguagem *Python* foram utilizadas para análise de dados, processamento de linguagem natural, agrupamento (*clustering)* e visualização.

O acesso aos dados foi realizado por meio do *Google Drive*, montado diretamente no ambiente *Colab*. A estruturação e manipulação dos dados foram realizadas com as bibliotecas *Python*: *pandas*: manipulação e análise de dados em estruturas tabulares. *numpy*: operações matemáticas e manipulação de arrays. *matplotlib*, *seaborn, plotly:* criação de gráficos e visualizações interativas. *scikit-learn (sklearn):* ferramentas de aprendizado de máquina, incluindo vetorização de textos, cálculo de similaridade, algoritmos de *clustering* e técnicas de redução de dimensionalidade. *re, string*: pré-processamento textual e limpeza de dados. *networkx*: construção e análise de grafos. *sentence-transformers*: geração de *embeddings* semânticos para representações vetoriais de textos. *scipy.spatial.distance*: cálculo de distâncias e similaridades entre vetores. *yellowbrick*: visualização de métricas e diagnósticos de modelos de *machine learning*.

Nesse contexto, o uso integrado dessas bibliotecas permite produzir um *pipeline* robusto para extração de conhecimento a partir dos textos, desde o pré-processamento até a visualização e avaliação preditiva.

Revista ARACÊ
ISSN: 2358-2472

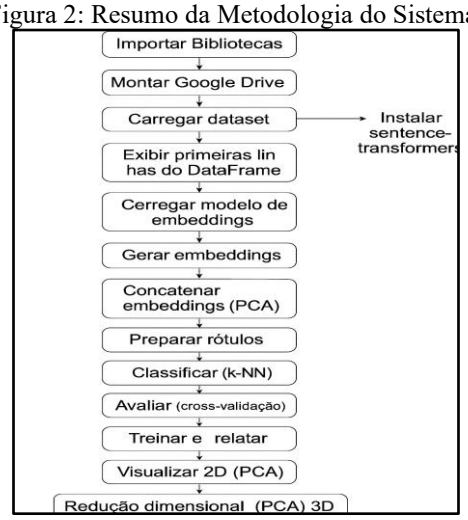
A montagem do ambiente e carregamento de dados no *Google Drive* foi montado para facilitar o acesso a arquivos remotos, e o *dataset* "*Multilingual_classification_dataset.xlsx*" onde foi carregado em um *DataFrame*

3.3 PROCEDIMENTOS DE EXECUÇÃO

As bibliotecas necessárias foram instaladas na célula de configuração inicial com os seguintes comandos: *!pip install pandas openpyxl numpy plotly ipykernel nbformat; !pip install -U sentence-transformers.* Essa configuração proporcionou um ambiente robusto e flexível para o desenvolvimento e execução das análises propostas ao longo do estudo.

Foram utilizadas técnicas TF-*IDF* (*Term Frequency-Inverse Document Frequency*), similaridade do cosseno e o algoritmo de agrupamento *K-Means*, bem como, vetorização e técnicas de redução de dimensionalidade *PCA (Principal Component Analysis) e t-SNE (t-Distributed Stochastic Neighbor Embedding*), para execução do dataset proposto.Na Figura 02 e descrito a sequência de elaboração e execução do sistemas (Figura 02).

Figura 2: Resumo da Metodologia do Sistema.



Fonte: Autores (2025).

Ressalta-se que a visualização de dados em 2D e 3D, utilizando técnicas como t-*SNE (t-Distributed Stochastic Neighbor Embedding)* ou *PCA (Principal Component Analysis),* desempenha um papel de suma importância na análise de vetores de alta dimensão, especialmente em tarefas como processamento de linguagem natural, aprendizado de máquina e reconhecimento de padrões, como no caso em tela. Essas técnicas reduzem a dimensionalidade dos dados complexos, como dados de documentos de patentes.

Revista ARACÊ
ISSN: 2358-2472

3.4 PROCEDIMENTOS METODOLÓGICOS

Nesta etapa, são detalhadas as etapas de coleta de dados (Por meio de um *dataset* envolvendo patentes), pré-processamento, classificação (*Data Mining*), interpretação e avaliação dos resultados.

Para os fins desta análise, que se concentra na similaridade textual, foram considerados inicialmente de forma exclusiva os resumos em inglês (coluna *'abs_English'*), uma vez que as versões em outros idiomas não eram relevantes para o escopo do processamento linguístico adotado. Os dados estão distribuídos em 6 classes/tipos de patentes, sendo que, os dados estão em sequência, abaixo os termos em português seguidos dos correspondentes em chinês: *1. Sensor (传感器), 2. Equipamento de carregamento (充电设备), 3. Cosméticos (化妆品), 4. Monitor (显示器), 5. Computador (计算机), 6. Alimento (食品).*

Por conseguinte, na vetorização do texto, foram consideradas as colunas com os seguintes resumos: *'abs_Chinese', 'abs_Original' e 'Class'*, que foram inicialmente removidas a fim de simplificar o *dataset* e manter o foco nos dados textuais em inglês (coluna *'abs_English'*).

Essa etapa garantiu a preparação adequada dos dados para as fases subsequentes de vetorização, cálculo de similaridade e modelagem por agrupamento, assegurando consistência e relevância dos textos analisados.

Para apoio computacional foram utilizadas ferramentas de Inteligência Artificial (IA), que foram empregadas para otimizar tanto a geração dos *scripts* quanto a elaboração textual do artigo, sempre sob supervisão e revisão. Destacam-se, nesse contexto, o uso do *Gemini*, da *Google*, na formulação dos códigos em *Python*, e do *ChatGPT*, da *OpenAI*, como suporte na redação técnica e metodológica.

4 RESULTADOS E DISCUSSÕES

A identificação de padrões semânticos e estruturais entre documentos técnicos é de suma importância para explorar a estrutura relacional entre diferentes patentes e identificar tendências de agrupamento temático.

A partir de 2010, a área de Processamento de Linguagem Natural (NLP) passou por avanços significativos com o desenvolvimento de técnicas de deep learning, sobretudo em tarefas como sumarização de texto, classificação de tópicos, análise de sentimentos e sintetização de voz. Com o surgimento de modelos instrucionais como o ChatGPT e o Gemini, especialmente a partir de 2022, observou-se uma transformação expressiva na qualidade dessas tarefas e um impacto notável da inteligência artificial generativa em múltiplos setores, acompanhando o progresso tecnológico e

Revista ARACÊ
ISSN: 2358-2472

contribuindo para o aumento da produtividade em diversas áreas (Morais *et al.,* 2025).

Já as patentes representam uma fonte de suma importância e de forma estratégica de conhecimento tecnológico, sendo sua correta classificação essencial para facilitar a recuperação e análise de informações relevantes em processos de inovação, por meio da análise em documentos de patentes. O uso de modelos baseados em engenharia de conhecimento permite melhorar significativamente a organização das patentes, tornando mais preciso o mapeamento de áreas tecnológicas e o apoio à tomada de decisão em instituições que atuam com propriedade intelectual no meio acadêmico e profissional (Wolski; Pizoni; Gonçalves, 2022).

4.1 ANÁLISE DE SIMILARIDADE (TF-IDF E COSSENO - *SENTENCE TRANSFORMERS*)

Conforme Souza; Semcovici; Pardo (2025) destacam-se a importância de implementar computacionalmente um modelo no sentido de explicitar dados. Demonstrando ser uma alternativa bastante interessante frente às propostas neurais atuais da Inteligência Artificial, em que há dificuldade de se obter explicabilidade das correlações semânticas, que podem produzir equívocos e "alucinações" (informações incorretas, inventadas ou sem base nos dados reais de treinamento).

A similaridade textual inicial foi calculada a partir da matriz TF-IDF, combinada com a similaridade do cosseno para mensurar a proximidade entre os documentos. Entretanto, o TF-IDF apresenta limitações relevantes: ele não captura aspectos semânticos mais profundos, como sinônimos ou relações contextuais entre palavras. Dessa forma, textos que expressam ideias semelhantes, mas com vocabulário distinto, tendem a ser classificados como pouco semelhantes, evidenciando a necessidade de métodos mais sofisticados, como *embeddings*.

Para superar as limitações do *TF-IDF*, foram utilizados *embeddings* gerados pelo modelo *Sentence Transformers (paraphrase-MiniLM-L6-v2)*, escolhido por equilibrar precisão e custo computacional. Os *embeddings* foram calculados para as colunas textuais do dataset, representando cada documento como um vetor denso em espaço de alta dimensionalidade.

4.2 USO DE EMBEDDINGS - GERAÇÃO DE VETORES DE REPRESENTAÇÕES SEMÂNTICAS

Avançando nas análises, por conseguinte, foi feito uso de embeddings semânticos avançados, com objetivo de substituir e complementar a vetorização *TF-IDF* por modelos de representação vetorial baseados em aprendizado profundo, como *Sentence-BERT/SBERT*, que pode capturar melhor as nuances semânticas dos textos e melhorar os resultados de similaridade e agrupamento. Nessa seara, observa-se que os *embeddings* que ficam próximos uns dos outros, têm valores próximos e

Revista ARACÊ
ISSN: 2358-2472

populações comparáveis, e podem ser considerados semelhantes. Usando *embeddings*, um algoritmo pode sugerir uma patente relevante, encontrar dados semelhantes ou identificar quais palavras provavelmente serão usadas juntas ou que são semelhantes umas às outras, como nos modelos de linguagem. Nesse caso, o sistema protótipo criado utilizou-se desses mecanismos (Marques; Gonçalves, 2024).

Por conseguinte, nesse contexto, os modelos são criados através do processo de treinamento baseado em informações de concorrência entre palavras. Constitui-se, assim, em uma ferramenta emergente para o processamento de linguagem natural sendo utilizada em uma ampla variedade de tarefas de processamento de idiomas. Sua utilidade decorre da capacidade de codificar relacionamentos de palavras no espaço vetorial. As aplicações variam desde componentes em sistemas de processamento de linguagem natural até ferramentas para análise linguística no estudo de linguagem e literatura (Souza, 2020 *apud* Heimerl; Gleicher, 2018).

Logo, em observância ao caso concreto, as colunas textuais foram convertidas em vetores de *embeddings*, preservando nuances semânticas fundamentais, especialmente importantes em *datasets* multilíngues (Figura 03 e 04). Os *embeddings* de cada coluna foram concatenados linha a linha, resultando em uma matriz densa de representações vetoriais (Figura 03).

Figura 3: Vetores Numéricos Gerados de Documentos de Patentes.

```
                                abs_English_embedding  \
0  [-0.0014463565312325954, -0.050861652940511703...
1  [-0.0051694829016923904, -0.05041619390249252,...
2  [0.013693631626665592, -0.05860981345176697, -...
3  [-0.02306394837796688, -0.03513801842927933, -...
4  [0.026903128251433372, 0.005832806695252657, -...

                               abs_Original_embedding  \
0  [0.019066711887717247, -0.007524144370108843, ...
1  [0.013877492398023605, -0.018452433869242668, ...
2  [0.019952455535531044, -0.004950828850269318, ...
3  [0.03973731771111488, -0.04726533591747284, -0...
4  [0.04324410483241081, 0.028452418744564056, 0....

                                     Class_embedding
0  [0.0018902403535321355, 0.05549944192171097, -...
1  [0.0018902403535321355, 0.05549944192171097, -...
2  [0.0018902403535321355, 0.05549944192171097, -...
3  [0.0018902403535321355, 0.05549944192171097, -...
4  [0.0018902403535321355, 0.05549944192171097, -...
```

Fonte: Autores (2025).

Logo, os dados foram vetorizados conforme Figura 03. Esses vetores numéricos densos gerados pelo *s-bert* são úteis em várias tarefas de processamento de linguagem natural, como recuperação de informações, classificação de texto, agrupamento de documentos e muito mais. Eles permitem comparar semanticamente as sentenças e calcular a similaridade entre elas de forma eficaz, o que é valioso em muitas aplicações (Marques; Gonçalves, 2024).

Revista ARACÊ
ISSN: 2358-2472

4.2.1 Embedding de Texto com Sentence Transformers

O *SentenceTransformers* é um *framework Python* para *embeddings* de sentenças, texto (Figura 04). Já o *DataFrame* é uma estrutura de dados que organiza os dados em uma tabela bidimensional de linhas e colunas, como uma planilha. Os *DataFrame*s são uma das estruturas de dados mais comuns na análise de dados moderna, pois são uma maneira flexível e intuitiva de armazenar e trabalhar com dados. Esses atributos foram utilizados no sistema, para melhor execução e alcance de seu objetivo (Figura 03 e 04).

Figura 4: Códigos Gerados para Vetorização Numéricos dos Documentos de Patentes.

```
!pip install -q sentence-transformers

import pandas as pd
import numpy as np
from sentence_transformers import SentenceTransformer, util
from scipy.stats import pearsonr

[ ] # Load model locally
from sentence_transformers import SentenceTransformer
model = SentenceTransformer("sentence-transformers/all-mpnet-base-v2")

import pandas as pd # dataframe manipulation
import numpy as np # linear algebra

# data visualization
import matplotlib.pyplot as plt
import matplotlib.cm as cm
import plotly.express as px
import plotly.graph_objects as go
import seaborn as sns
import shap

# sklearn
from sklearn.cluster import KMeans
from sklearn.preprocessing import PowerTransformer, OrdinalEncoder
from sklearn.pipeline import Pipeline
from sklearn.manifold import TSNE
from sklearn.metrics import silhouette_score, silhouette_samples, accuracy_score, classification_report
```

Fonte: Autores (2025).

Risch e Krestel (2019), ao analisarem documentos de patentes, observaram que o uso de embeddings de palavras treinadas especificamente para esse domínio resultou em um aumento expressivo na precisão das classificações automáticas. Os autores demonstram que tais *embeddings* são eficazes na captura dos padrões linguísticos técnicos característicos das patentes.

Dessa forma, a principal vantagem dessa abordagem é a capacidade de capturar relações semânticas de forma contextualizada, o que possibilita a identificação de semelhanças mais profundas entre documentos, mesmo quando o vocabulário difere. No contexto multilíngue, isso foi particularmente relevante, pois permitiu mapear textos em diferentes idiomas em um mesmo espaço vetorial.

4.2.2 Redução de Dimensionalidade e Visualização

De acordo com Krestel *et al.* (2021), os métodos mais recentes de representação textual baseiam-se em *embeddings* de palavras contextuais, como o *Bert*, que é construído sobre a arquitetura de transformadores. Esses modelos aprendem representações específicas para cada palavra com base

Revista ARACÊ
ISSN: 2358-2472

em seu contexto na frase, e têm se mostrado altamente eficazes em tarefas como classificação de patentes. Além disso, modelos como o GPT vêm sendo explorados para a geração de reivindicações de patentes.

Nessa linha, no caso concreto, para melhor viabilizar a visualização, foram aplicadas técnicas de *PCA (Principal Component Analysis)* e *t-SNE (t-Distributed Stochastic Neighbor Embedding)*.

4.2.3 Agrupamento (Clustering) com K-Means

O algoritmo *K-Means* foi aplicado para segmentar os documentos em *K clusters* previamente definidos. Embora a análise do conteúdo interno dos grupos não tenha sido detalhada no *notebook,* é possível inferir que os clusters refletem proximidade semântica entre os documentos.

Por conseguinte, a coerência dos agrupamentos pôde ser validada pela comparação com as visualizações de PCA e t-SNE, que sugeriram consistência parcial. Em alguns casos, os clusters revelaram divisões temáticas claras, enquanto em outros houve dispersão interna, indicando sobreposição de conteúdos ou ausência de fronteiras rígidas entre tópicos.

De acordo com autores como Reymond e Dematraz (2014), a análise de redes pode ser aplicada à exploração de dados de patentes, permitindo que conexões ocultas entre inventores, empresas e tecnologias sejam reveladas e visualizadas de forma mais eficiente, facilitando decisões estratégicas no âmbito da inteligência competitiva. Segundo os pesquisadores, o uso de redes na análise de patentes ajuda a identificar padrões e relações entre classificações internacionais de patentes (IPC) e países, destacando a disseminação tecnológica ao longo do tempo e a especialização em determinados setores.

Não obstante, ressalta-se que a automatização do processo de classificação e visualização de dados de patentes pode atuar como uma ferramenta eficaz de busca e recuperação de informações, sendo útil tanto na identificação de oportunidades quanto na prevenção de infrações às leis de propriedade intelectual (Martins; Francisco; Farias, 2021).

4.2.4 Classificação com K-Nearest Neighbors (KNN)

A etapa de classificação teve como objetivo prever a variável *'Class'* a partir dos *embeddings*. Para isso, utilizou-se o algoritmo *KNeighborsClassifier*, adequado para dados vetoriais e com implementação eficiente.

A avaliação do modelo foi conduzida por meio de divisão treino/teste e validação cruzada, assegurando maior robustez na medição do desempenho.

Revista ARACÊ
ISSN: 2358-2472

O relatório de classificação indicou uma acurácia geral de 95%. Em termos de métricas por classe, algumas apresentaram alta precisão e recall, sugerindo que os documentos dessas categorias possuem características textuais bem definidas. Outras, no entanto, mostraram baixo desempenho, com desequilíbrio entre precisão e recall. Esse comportamento indica que, para algumas classes, o modelo produziu mais falsos positivos, enquanto para outras houve prevalência de falsos negativos.

O desempenho heterogêneo pode estar relacionado à distribuição desigual das classes no dataset, bem como à proximidade semântica entre determinadas categorias, que dificulta a distinção pelo KNN.

O *k-NN* foi escolhido em detrimento de modelos mais complexos (como SVM ou redes neurais) por sua capacidade de refletir diretamente a estrutura semântica dos dados vetoriais, sem requerer treinamento intensivo.

4.3 REDUÇÃO DE DIMENSIONALIDADE - CODIFICAÇÃO DOS DADOS E RÓTULOS

Nos últimos anos, modelos de linguagem baseados em *deep learning* têm revolucionado o campo do Processamento de Linguagem Natural (PLN) ao possibilitar a geração de *embeddings* textuais, logo, vetores densos são capazes de representar, de forma numérica, o significado semântico de palavras, frases e documentos. Assim, na sequência, são apresentados os gráficos resultantes do agrupamento dos documentos por similaridade, obtidos a partir das técnicas anteriormente descritas, utilizando, em todas as análises, o mesmo *dataset* contendo os resumos das patentes.

4.3.1 Análise de Componentes Principais (PCA)

O gráfico de dispersão 2D gerado usando PCA (Análise de Componentes Principais) exibe os dados projetados na forma plana. Observa-se a formação de alguns clusters, indicando que a PCA conseguiu capturar parte da estrutura de agrupamento das classes nos dados de *embedding* (Figura 05).

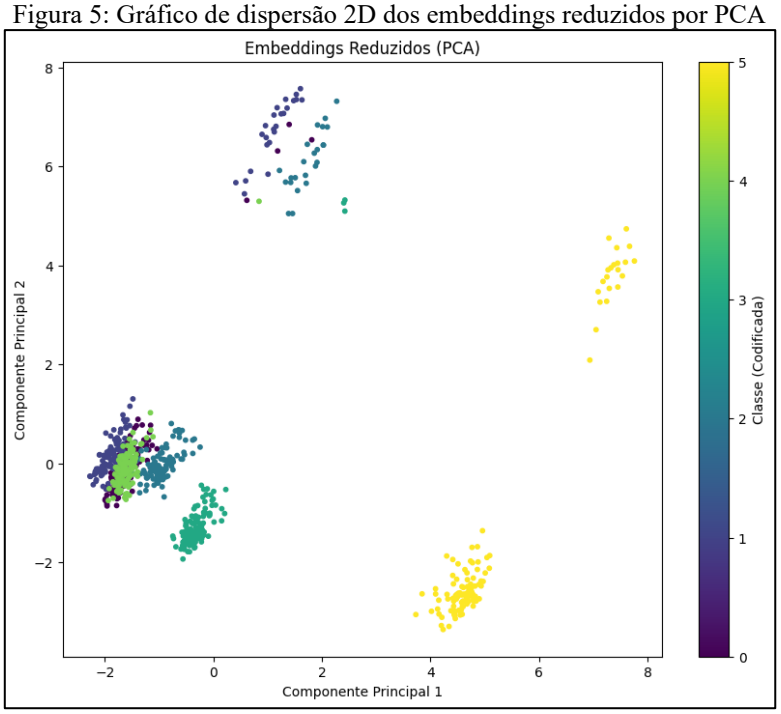
Dessa forma, a descrição dos gráficos de redução de dimensionalidade do gráfico de dispersão 2D gerado usando PCA destaca que há certa sobreposição entre os clusters, sugerindo que a separabilidade das classes não é perfeita apenas com as duas primeiras componentes principais (Figura 05).

Nesse diapasão, observa-se que modelos de *embeddings* pré-treinados em grandes corpora multilíngues, como os da família `*sentence-transformers*`, oferecem a capacidade de gerar representações vetoriais que são semanticamente significativas e comparáveis entre diferentes

Revista ARACÊ
ISSN: 2358-2472

idiomas. Esta propriedade é particularmente valiosa para tarefas que envolvem dados textuais de diversas origens linguísticas (Figura 05).

Figura 5: Gráfico de dispersão 2D dos embeddings reduzidos por PCA



Fonte: Autores (2025).

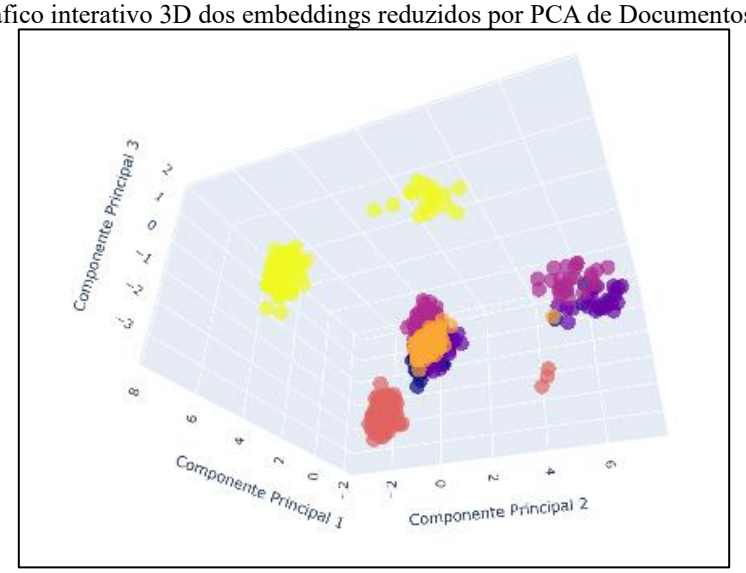
Nessa esteira, os *embeddings* gerados a partir de colunas textuais de um *dataset* multilíngue podem ser utilizados como features para um modelo de classificação e como técnicas de redução de dimensionalidade, que podem auxiliar na análise exploratória e, potencialmente, melhorar o desempenho do modelo.

Como já destacado, com o grande volume de depósitos de patentes, torna-se essencial o uso de novas ferramentas computacionais, com intuito de apoiar a análise, classificação e visualização dessas informações em documentos de patentes (Figura 06). Malgrado, técnicas baseadas em Processamento de Linguagem Natural (PLN) e aprendizado de máquina vêm se mostrando promissoras para extrair conhecimento estruturado a partir de textos de patentes, facilitando a triagem tecnológica, a detecção de similaridades e o mapeamento de inovações (Figura 06).

A visualização interativa 3D gerada estende a representação para três dimensões, permitindo uma exploração mais completa da estrutura dos dados. Nesta visualização, os clusters se tornam ainda mais espacialmente separados, facilitando a distinção visual entre a maioria das classes (Figura 06).

Revista ARACÊ
ISSN: 2358-2472

Figura 6: Gráfico interativo 3D dos embeddings reduzidos por PCA de Documentos de Patentes.



Fonte: Autores (2025).

Nessa linha, uma vez que as palavras são representadas como vetores, podemos calcular a similaridade entre elas usando métricas como similaridade de cosseno, distância euclidiana ou correlação de *Pearson*. Estas métricas quantificam a proximidade entre os vetores de palavras (Figura 06).

O presente estudo empregou uma série de técnicas avançadas de Processamento de Linguagem Natural (PLN), aprendizado de máquina e visualização de dados com o objetivo de explorar a estrutura latente de um conjunto de dados multilíngues de classificação textual.

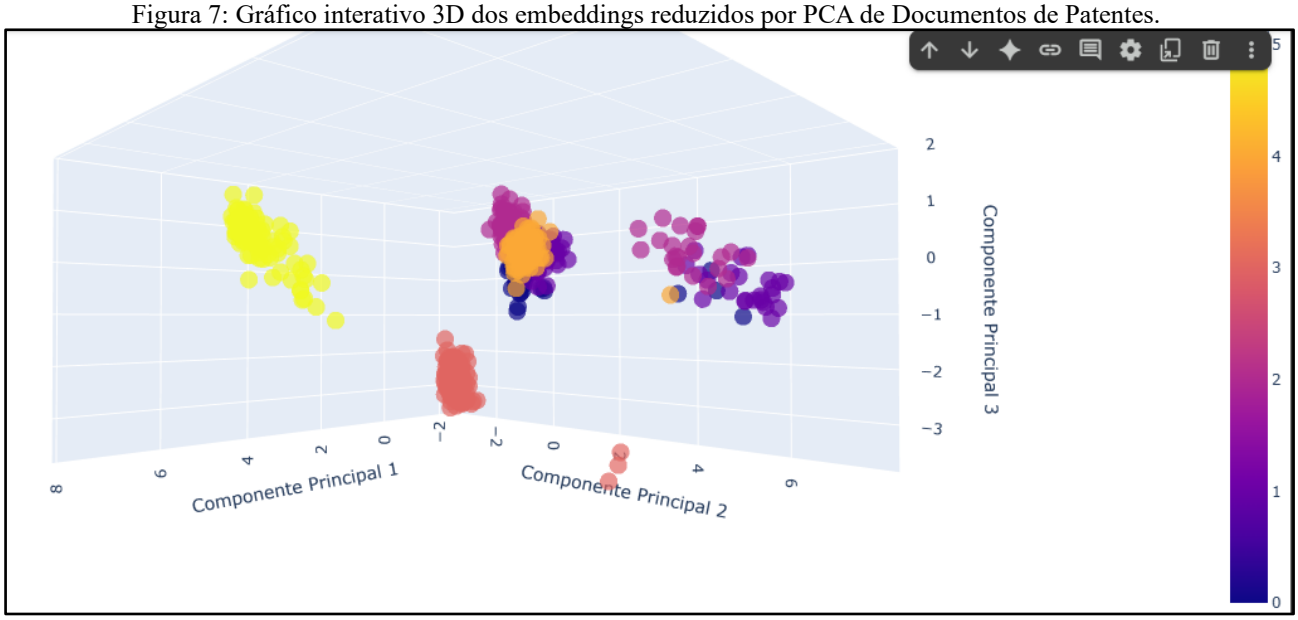
Foi utilizado o classificador *K-Nearest Neighbors (k-NN)* com k=3k = 3k=3, escolhido por sua simplicidade interpretativa e sua eficácia em problemas com representações vetoriais densas. O modelo foi avaliado com validação cruzada estratificada de *5 folds*, resultando em uma acurácia média consistente, refletindo a separabilidade dos embeddings reduzidos no espaço vetorial. Na etapa de teste, foram calculadas as métricas padrões (*acurácia, precisão, recall, F1-score)* por classe.

Nessa linha, o gráfico de dispersão dos embeddings reduzidos por PCA permite observar a distribuição dos documentos em um espaço bidimensional, capturando a maior variância dos dados e projetando vetores de alta dimensão em componentes principais.

Por conseguinte, nota-se a formação de agrupamentos de pontos coloridos que refletem as classes originais dos documentos, sendo a separação e a clareza desses agrupamentos um indicativo da eficácia do PCA em distinguir padrões. Contudo, a presença de sobreposição entre cores evidencia que certas classes compartilham semelhanças relevantes nas direções de maior variância, o que pode limitar a capacidade do método em representar diferenças mais sutis entre os grupos (Figura 07).

Revista ARACÊ
ISSN: 2358-2472

Figura 7: Gráfico interativo 3D dos embeddings reduzidos por PCA de Documentos de Patentes.



Fonte: Autores (2025).

Logo, nesse diapasão, esses resultados demonstram que, embora o PCA seja útil para oferecer uma primeira visualização global das relações entre as classes, sua limitação na captura de não linearidades pode comprometer a separação completa entre os documentos.

Nesse sentido, a comparação com técnicas alternativas de redução de dimensionalidade, como o t-SNE, torna-se estratégica para compreender melhor a estrutura dos dados e verificar se padrões mais complexos podem ser revelados.

Assim, o gráfico contribui para discutir a adequação das técnicas empregadas, fornecendo evidências tanto da utilidade quanto das limitações do PCA na análise exploratória dos embeddings (Figura 07).

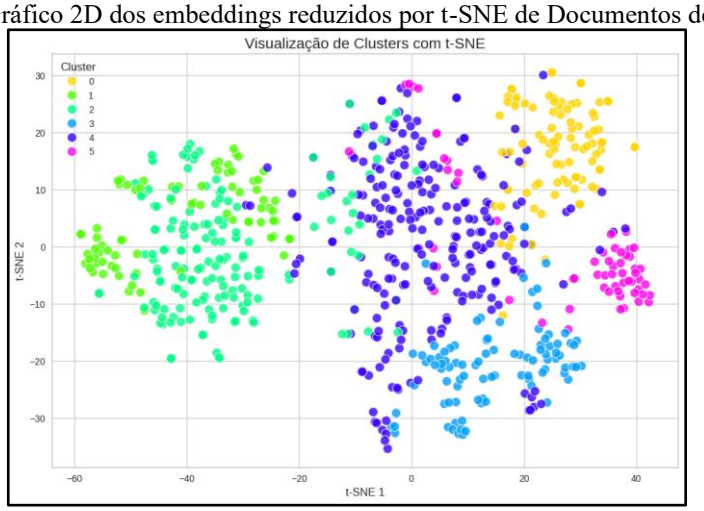
4.3.2 Técnica t-Distributed Stochastic Neighbor Embedding (t-SNE)

O gráfico de dispersão 2D gerado usando t-SNE (t-Distributed Stochastic Neighbor Embedding) mostra uma separação de clusters mais clara e compacta em comparação com o PCA 2D. O t-SNE é mais eficaz em preservar as distâncias locais entre os pontos de alta dimensão, resultando em agrupamentos visuais mais distintos para as diferentes classes. É possível identificar agrupamentos bem definidos para a maioria das classes, embora alguns pontos de classes diferentes possam estar próximos (Figura 08).

Na figura 08 abaixo, analisamos a dispersão dos pontos que representam as das 720 patentes da amostragem, com a demonstração das proximidades entre os pontos, sendo registrada as semelhanças entre as patentes por suas distâncias e relação entre os pontos.

Revista ARACÊ
ISSN: 2358-2472

Figura 8: Gráfico 2D dos embeddings reduzidos por t-SNE de Documentos de Patentes.



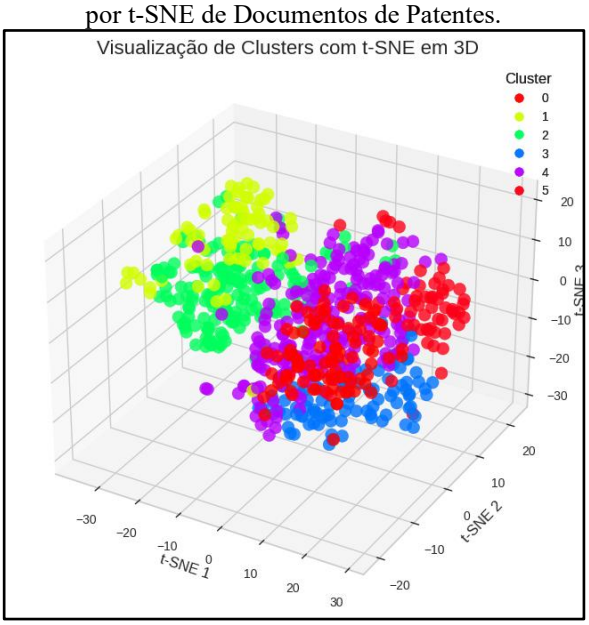
Fonte: Autores (2025).

A visualização interativa 3D gerada usando t-SNE estende a representação para três dimensões, permitindo uma exploração mais completa da estrutura dos dados. Nesta visualização, os clusters se tornam ainda mais espacialmente separados, facilitando a distinção visual entre a maioria das classes. A perspectiva 3D ajuda a revelar a complexidade da distribuição dos dados no espaço de embeddings e a eficácia do t-SNE em separar os agrupamentos (Figura 09).

O gráfico de dispersão dos *embeddings* reduzidos pelo t-SNE evidencia como os documentos se organizam no espaço bidimensional ao preservar, na medida do possível, as distâncias do espaço original de alta dimensão. Observa-se a formação de agrupamentos de pontos coloridos, representando as classes, cuja proximidade ou separação indica o grau de similaridade semântica entre os documentos.

Revista ARACÊ
ISSN: 2358-2472

Figura 9: Modelo de distribuição das patentes, após transformação em Vetores – Gráfico 3D dos embeddings reduzidos por t-SNE de Documentos de Patentes.



Fonte: Autores (2025).

A presença de clusters bem definidos sugere que os embeddings capturaram características relevantes, enquanto a sobreposição entre cores aponta para semelhanças semânticas entre classes distintas ou limitações introduzidas pelo processo de redução de dimensionalidade. Importa destacar que o t-SNE, por ser um método não linear e estocástico, pode gerar variações em diferentes execuções, ainda que a estrutura geral de agrupamentos se mantenha consistente (Figura 09).

Logo, esses resultados reforçam a utilidade do t-SNE como ferramenta exploratória para compreender a estrutura semântica dos embeddings e avaliar a separabilidade das classes de documentos. A clareza dos clusters indica a robustez dos embeddings para fins de classificação, ao passo que áreas de sobreposição sinalizam oportunidades de melhoria, seja pela adoção de técnicas mais sofisticadas de pré-processamento, seja pela utilização de modelos alternativos de representação textual.

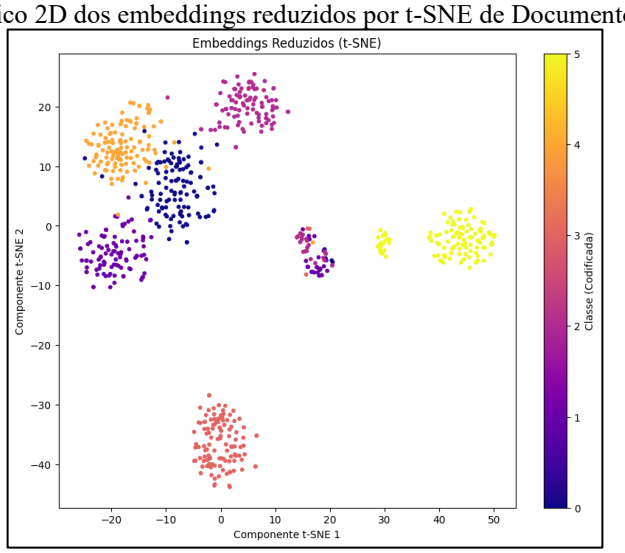
Quando comparado ao PCA, o t-SNE tende a oferecer maior capacidade de captura de padrões não lineares, fornecendo insights complementares que enriquecem a análise da qualidade dos embeddings e da adequação dos métodos de classificação adotados (Figura 10).

Ao comparar as representações em 2D e 3D, percebe-se que adicionar uma terceira dimensão, especialmente com t-SNE, frequentemente aprimora a separação visual dos clusters. Enquanto o PCA em 2D oferece uma visão geral da variância principal, o t-SNE, tanto em 2D quanto em 3D, tende a destacar a estrutura local e a formação de agrupamentos, com a versão 3D proporcionando uma perspectiva mais rica e tridimensional da distribuição das classes no espaço de embeddings.

Revista ARACÊ
ISSN: 2358-2472

O uso de gráficos na análise da similaridade de padrões nas patentes, não apenas agiliza o processo de análise, mas também aprimora a interpretação e a comunicação dos resultados, contribuindo significativamente para o avanço do conhecimento e a tomada de decisões fundamentadas (Figura 10).

Figura 10: Gráfico 2D dos embeddings reduzidos por t-SNE de Documentos de Patentes.



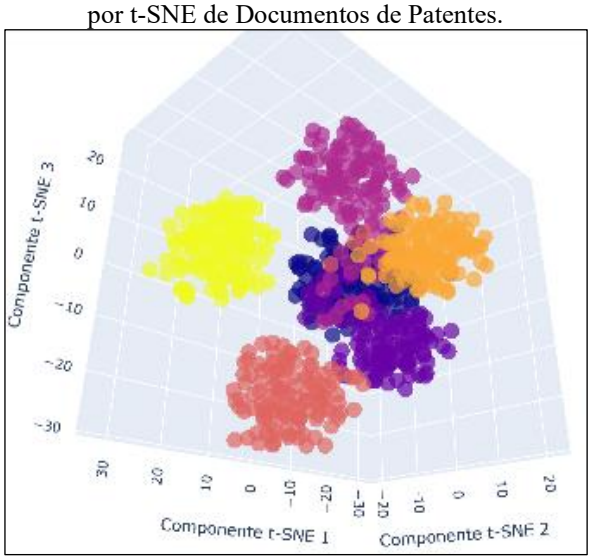
Fonte: Autores (2025).

Nessa toada, além disso, os gráficos facilitam a comunicação de resultados, tornando a informação mais acessível e compreensível para uma ampla gama de audiências. Dessa forma, não obstante, o uso de gráficos na análise da similaridade de padrões nas patentes, não apenas agiliza o processo de análise, mas também aprimora a interpretação e a comunicação dos resultados, contribuindo significativamente para o avanço do conhecimento, mormente na tomada de decisões fundamentadas (Marques; Gonçalves, 2024).

Os resultados da classificação e as visualizações em 2D e 3D demonstram o potencial dos embeddings textuais combinados com a redução de dimensionalidade para capturar padrões discriminativos em dados multilíngues, sugerindo uma abordagem promissora para tarefas de classificação em cenários complexos onde a representação semântica dos dados textuais é crucial (Figura 11).

Revista ARACÊ
ISSN: 2358-2472

Figura 11: Modelo de distribuição das patentes, após transformação em Vetores – Gráfico 3D dos embeddings reduzidos por t-SNE de Documentos de Patentes.



Fonte: Autores (2025).

Dessa forma, por meio das visualizações bidimensionais e tridimensionais dos *embeddings* reduzidos — geradas com o auxílio das bibliotecas matplotlib e *plotly.express* — foi possível representar visualmente as classes codificadas por meio da coloração dos pontos nos gráficos. As visualizações indicam uma separação parcial entre as classes, com a formação de *clusters* bem definidos em algumas categorias, evidenciando a existência de estrutura latente nos dados textuais.

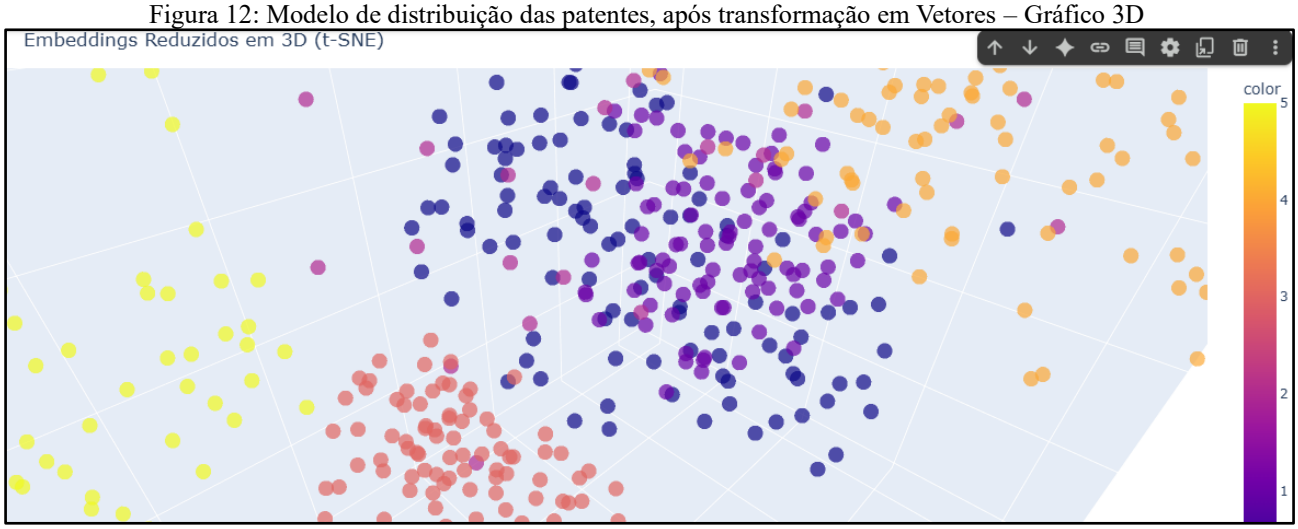
O desempenho do classificador *k-NN (k-Nearest Neighbors*) foi avaliado por meio da acurácia em validação cruzada e da análise do relatório de classificação sobre o conjunto de testes. Os resultados obtidos demonstram a viabilidade da abordagem proposta, indicando um desempenho promissor na tarefa de classificação, ainda que haja margem para aprimoramento na distinção entre classes específicas.

Os achados corroboram a hipótese de que *embeddings* textuais constituem uma representação eficaz para dados textuais em contextos multilíngues, sobretudo quando empregados com modelos da família *sentence-transformers*. A capacidade desses modelos de gerar representações vetoriais semanticamente consistentes entre diferentes idiomas revela-se um elemento central para o êxito da abordagem.

A aplicação de técnicas de redução de dimensionalidade, com destaque para o t-SNE, mostrou-se essencial para a compreensão da estrutura intrínseca dos dados no espaço vetorial, permitindo a visualização de padrões e da separabilidade entre classes que, de outro modo, permaneceram ocultos em espaços de alta dimensão (Figura 12).

Revista ARACÊ
ISSN: 2358-2472

Figura 12: Modelo de distribuição das patentes, após transformação em Vetores – Gráfico 3D



Fonte: Autores (2025).

Apesar da simplicidade inerente ao *k-NN*, sua aplicação no espaço de *embeddings* reduzido fornece uma linha de base robusta para a avaliação da "classificabilidade" dos dados. Destaca-se, por fim, que a utilização de classificadores mais sofisticados — como *Support Vector Machines* (SVMs) ou redes neurais profundas — poderá conduzir a melhorias adicionais de desempenho, especialmente em tarefas que exigem maior capacidade de modelagem não linear.

Pode-se inferir que a visualizações 2D e 3D com PCA e t-SNE, demonstrou por meio de estrutura vetorial dos dados, três técnicas de visualização: (01) PCA 2D: Revelou agrupamentos razoavelmente bem definidos, indicando que parte da separação entre classes pode ser explicada linearmente. (02) PCA 3D (interativo com Plotly): Facilitou a inspeção interativa da distribuição de classes, útil para identificar padrões de sobreposição. (03) t-SNE 2D e 3D: Como técnica não linear, o t-SNE evidenciou agrupamentos mais coesos entre classes semanticamente semelhantes, confirmando a existência de estrutura de similaridade latente entre documentos.

Dessa forma, os resultados obtidos demonstram que a concatenação de dados objeto de criação de *embeddings*, que são fruto de diferentes campos textuais proporcionou uma representação informacional mais rica, cuja estrutura latente foi evidenciada por meio das técnicas de redução de dimensionalidade (PCA e t-SNE) e visualizações em 2D e 3D, podendo identificar padrões e similaridades entres os diferentes documentos.

Outrossim, dessa mesma forma, o classificador k-NN apresentou desempenho satisfatório, validando a separabilidade das classes mesmo após redução da dimensionalidade, o que indica que as representações extraídas preservaram características discriminativas relevantes. A aplicação de técnicas de redução de dimensionalidade, como PCA e t-SNE, não apenas viabilizou a visualização

Revista ARACÊ
ISSN: 2358-2472

da estrutura latente do espaço semântico das patentes, mas também revelou agrupamentos coerentes entre documentos com similaridade técnica.

Logo, tanto o PCA quanto o t-SNE foram aplicados para reduzir a dimensionalidade dos *embeddings* textuais e visualizar a estrutura do *dataset multilíngue.* O PCA em 2D mostrou alguns agrupamentos, mas com sobreposição. O t-SNE, tanto em 2D quanto em 3D, demonstrou ser mais eficaz em revelar clusters distintos e bem separados para as diferentes classes, destacando sua utilidade na visualização da estrutura local dos dados de alta dimensão e na identificação da separabilidade das classes. Por fim, destaca-se que a visualização em 3D com t-SNE proporcionou a maior clareza na distinção entre os grupos.

4.3.3 Isolamento e Análise Individual de Elementos

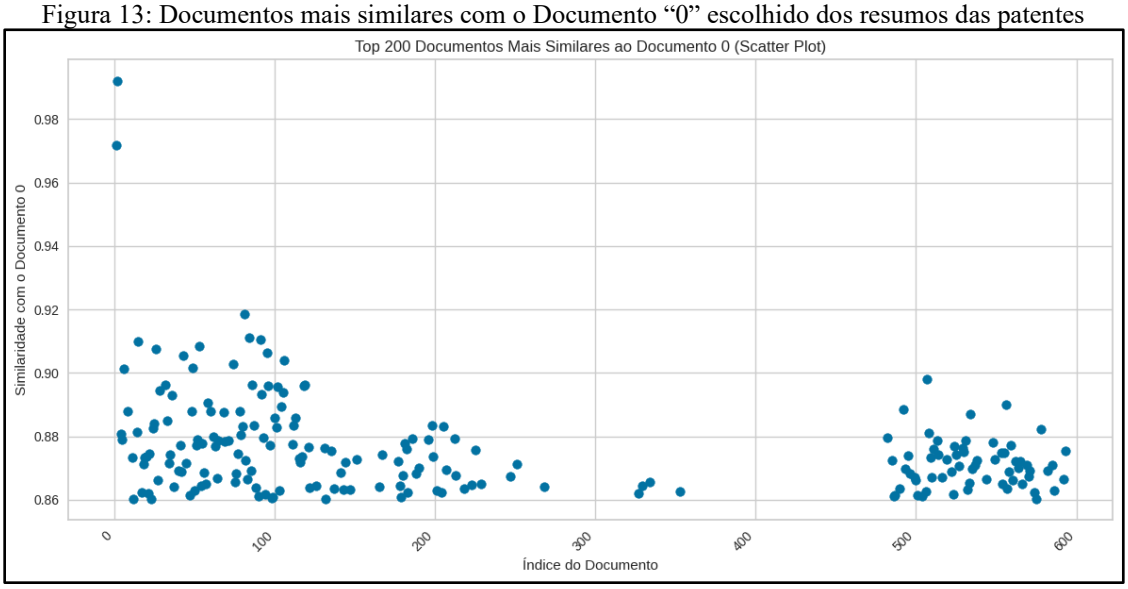
Num contexto, de forma mais singular, e foco em documentos específicos, o gráfico de dispersão dos 200 documentos mais similares ao documento de índice 0 evidencia como a métrica de similaridade de cosseno captura relações semânticas entre os textos. Nota-se que os documentos mais próximos semanticamente ao documento 0 atingem valores de similaridade superiores a 0,95, destacando-se logo no início do eixo X.

A maioria dos pontos, entretanto, concentra-se entre as faixas de 0,86 e 0,92, o que demonstra a existência de um núcleo consistente de documentos com proximidade temática significativa, ainda que não idêntica.

A dispersão ao longo do eixo X revela que a similaridade não está relacionada à posição dos documentos no dataset, mas ao conteúdo capturado pelos embeddings. Além disso, a distribuição em diferentes patamares de similaridade pode indicar a presença de subgrupos com distintos níveis de relação semântica com o documento 0.

Revista ARACÊ
ISSN: 2358-2472

Figura 13: Documentos mais similares com o Documento “0” escolhido dos resumos das patentes



Fonte: Autores (2025).

Dessa forma, esses resultados reforçam a utilidade da visualização na identificação dos documentos mais relevantes para estudos comparativos, permitindo distinguir não apenas os casos de maior proximidade, mas também os limites de similaridade dentro do grupo analisado. Não obstante, no entanto, a interpretação substantiva das relações exige a leitura dos resumos ou conteúdos originais dos documentos, uma vez que a métrica numérica apenas indica a intensidade da proximidade semântica.

Assim, a análise qualitativa desses documentos poderia revelar padrões lexicais, temáticos ou estruturais que fundamentam a similaridade observada, oferecendo subsídios adicionais para discussões sobre agrupamento semântico e consistência dos *embeddings* utilizados.

4.4 DISCUSSÃO GERAL E ACHADOS DOS RESULTADOS

Diferente do artigo *Descobrindo conexões e similaridades em textos de patentes: processamento de linguagem natural e visualização interativa”* que analisa a aplicação de técnicas de Processamento de Linguagem Natural (PLN), aprendizado de máquina e visualização de dados na identificação de similaridades entre documentos de patentes multilíngues, onde os autores destacam que as representações visuais em grafos e redes permitem compreender melhor as conexões entre tecnologias e revelar tendências emergentes, contribuindo para processos decisórios mais estratégicos no campo da propriedade intelectual. Nos gráficos bidimensionais e tridimensionais, observou-se a formação de agrupamentos dispersos, sugerindo a presença de padrões latentes nos dados. Quando os pontos foram coloridos segundo clusters/classes, constatou-se que alguns grupos apresentaram clara separação, enquanto outros mostraram sobreposição, refletindo a complexidade semântica dos

Revista ARACÊ
ISSN: 2358-2472

documentos.

Em termos comparativos, o PCA preservou bem a estrutura global dos dados, mas não evidenciou claramente os agrupamentos locais. O t-SNE, por sua vez, destacou com mais nitidez a separação de clusters, ainda que com maior sensibilidade a hiperparâmetros e a escalas locais. Essa diferença sugere que a estrutura dos dados possui tanto dimensões globais de variação quanto padrões locais mais sutis. A análise integrada dos dados nas diferentes etapas do processo de execução permite algumas sínteses importantes:

a) O pré-processamento contribuiu para reduzir ruídos e padronizar os textos, bem como, o TF-IDF forneceu uma análise inicial útil, mas insuficiente para capturar a riqueza semântica.
b) Os *embeddings* representaram uma evolução significativa, permitindo identificar relações mais profundas entre os documentos. Na mesma linha, a redução de dimensionalidade e as visualizações forneceram informações visuais sobre a estrutura dos dados, revelando agrupamentos coerentes, mas também sobreposições.
c) O K-Means evidenciou padrões latentes, embora com variação na coerência dos clusters. Já o KNN apresentou resultados razoáveis, mas seu desempenho foi limitado em classes mais próximas semanticamente ou com menor representatividade no dataset.

Por conseguinte, o estudo demonstra como técnicas de Processamento de Linguagem Natural (PLN), Inteligência Artificial (IA) e visualização interativa podem ser aplicadas na análise de documentos de patentes, contribuindo para a organização e interpretação de grandes volumes de informações tecnológicas.

Nessa linha, o estudo evidencia que a utilização de algoritmos de aprendizado de máquina, vetorização, grafos e análises semânticas possibilita identificar similaridades entre patentes, tendências emergentes e conexões estratégicas entre diferentes áreas do conhecimento, auxiliando na prospecção tecnológica, na inovação e na tomada de decisões. Bem como, nessa toada, observa-se que a pesquisa contribui para o avanço de sistemas inteligentes de busca patentária, reduzindo redundâncias, otimizando análises de anterioridade e fortalecendo a gestão do conhecimento tecnológico em ambientes acadêmicos e corporativos altamente dinâmicos.

Por fim, os resultados respondem diretamente ao objetivo de avaliar a similaridade, a estrutura e a classificação de documentos por meio de técnicas de PLN e aprendizado de máquina, demonstrando as potencialidades e as restrições de cada método.

Revista ARACÊ
ISSN: 2358-2472

5 CONCLUSÃO

As visualizações geradas forneceram dados e informações *valiosas e* significativas sobre a estrutura de similaridade do conjunto de dados, contribuindo para uma compreensão mais aprofundada das relações semânticas entre os documentos e possibilitando a identificação de clusters temáticos com maior precisão.

Nessa linha, a análise de similaridade semântica entre textos de patentes, por meio de técnicas de Processamento de Linguagem Natural para melhor visualização de dados, revelou-se uma abordagem eficaz para compreender padrões ocultos em grandes volumes de documentos técnicos. A aplicação de vetorização textual com TF-IDF, combinada com algoritmos de redução de dimensionalidade como PCA e t-SNE, permitiu estruturar semanticamente os resumos de patentes multilíngues, evidenciando agrupamentos temáticos e áreas tecnológicas críticas.

As visualizações interativas geradas a partir dos dados e informações demonstraram grande potencial como ferramentas estratégicas para pesquisadores e para gestão da informação tecnológica. Elas oferecem suporte valioso para examinadores, pesquisadores e formuladores de políticas públicas, ao facilitar a identificação de tendências emergentes, desafios específicos no processo de concessão de patentes e padrões linguísticos característicos de diferentes domínios tecnológicos.

Dessa forma, o estudo contribui significativamente para o avanço das investigações sobre linguagem técnica e científica em documentos de patentes, ao propor uma metodologia robusta e visualmente acessível para a análise de documentos complexos. Logo, ao integrar técnicas computacionais com uma abordagem visual, o artigo reforça a importância da interdisciplinaridade na construção de soluções concretas para os desafios da busca e análise de patentes.

Para pesquisas futuras, recomenda-se ampliar o escopo dos dados analisados, incorporando outras variáveis que possam influenciar as decisões de patenteamento e um número maior de dados. Entre as limitações que podemos destacar nesse estudo incluem a dependência da qualidade dos embeddings gerados pelo modelo pré-treinado e a sensibilidade das técnicas de redução de dimensionalidade (particularmente t-SNE) aos parâmetros escolhidos.

AGRADECIMENTOS

Este trabalho foi realizado com o apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES).

Revista ARACÊ
ISSN: 2358-2472

REFERÊNCIAS

ALONSO-MARTÍNEZ, Daniel; GONZÁLEZ-ÁLVAREZ, Nuria; NIETO, Mariano. Does international patent collaboration have an effect on entrepreneurship? Journal of International Entrepreneurship, v. 19, p. 539–559, 2021.

ALVES, Renato Lourenço; SOUZA, Paulo Augusto Ramalho de; NEDER, Renato. Análise de Patentes Através de Redes Semânticas: A Inteligência Artificial no Agronegócio entre 2009 e 2018. XLVI Encontro da ANPAD - EnANPAD 2022, São Caetano do Sul, SP, 21-23 set. 2022.

BASSO, Fernanda G. Análise de similaridade em textos de patentes com Processamento de Linguagem Natural: representações vetoriais e redução de dimensionalidade – uma abordagem visual. 2020. Dissertação – Universidade de São Paulo, São Paulo, 2020.

BHATT, Priyanka C.; MISHRA, Durgesh K.; SHRIVASTAVA, Garima. Patent analysis-based technology innovation assessment with the lens of disruptive innovation theory: A case of blockchain technological trajectories. *Technological Forecasting and Social Change,* v. 198, 123905, 2023. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S0040162523005498. Acesso em: 5 abr. 2025.

BRITO, Ana Paula Damasceno; OZAKI, Adalton Masalu. Busca patentária: a chave do sucesso em projetos tecnológicos INOVA IFSP. CONICT. 2019.

CÂNDIDO, Rafael; GONÇALVES, Alexandre Leopoldo; LEMOS, Robson Rodrigues. Information Visualization to Support Idea Management. IEEE Latin America Transactions, vol. 20, n. 6, jun. 2022.

DAMO, Emerson. Mapeamento do desenvolvimento tecnológico na indústria por meio de documentos de patentes: análise da inovação na indústria automobilística. 2021. 106 f. Dissertação (Mestrado Profissional em Gestão e Tecnologia em Sistemas Produtivos) – Centro Estadual de Educação Tecnológica Paula Souza, São Paulo, 2021

DUARTE, José Mateus Rodrigues Farias; FILHO, Antonio Carlos de Sousa; GIRÃO, Mauro Vinícius Dutra. Nuvens de palavras auxiliando no aprendizado de Fisiologia Humana: relato de experiência. Revista de Educación en Biología, v. 26, n. 2, p. 24-38, 2023.

INSTITUTO NACIONAL DA PROPRIEDADE INDUSTRIAL. Governança. Ministério do Desenvolvimento, Indústria, Comércio e Serviços. Disponível em: https://www.gov.br/inpi/pt-br/governanca. Acesso em: 9 ago. 2025

KRESTEL, R., CHIKKAMATH, R., HEWEL, C., & RISCH, J. A survey on deep learning for patent analysis. World Patent Information, 65, 102035. 2021.

LEAL, Alfredo Julio; CORTESE, Tatiana; KNIESS, Cláudia Terezinha. Contribuição das informações patentárias na busca de tecnologias para reciclagem do resíduo de equipamento eletroeletrônico. Anais do ENGEMAUSP, 2015. Disponível em: https://engemausp.submissao.com.br/17/anais/arquivos/277.pdf. Acesso em: 1 jun. 2025.

Revista ARACÊ
ISSN: 2358-2472

Liu, W.; Zhang, Y.; Luo, X.; Cao, Y.; Gan, K.; Ye, F.; Tang, W.; Zhang, M. Patent transformation prediction: When a patent can be transformed. Information Processing & Management, [S.l.], v. 61, n. 6, p. 103872, nov. 2024. Disponível em: https://doi.org/10.1016/j.ipm.2024.103872. Acesso em: 4 jul. 2025.

MARQUES, Thiago Domingos; GONÇALVES, Alexandre Leopoldo. A importância de um sistema de organização de patentes por análise semântica: proposta de um protótipo. In: X ENPI, 2024, Fortaleza-CE. Anais do X ENPI. Fortaleza-CE: 2024. v. 10, n. 1, p. 01-06.

MARQUES, T. D.; GONÇALVES, A. L.; PAULINO, R. de C. R.; SOUZA, M. V. de; DANDOLINI, G. A. Descobrindo conexões e similaridades em textos de patentes: processamento de linguagem natural e visualização interativa. Revista Delos, [S. l.], v. 18, n. 69, p. e5912, 2025. DOI: 10.55905/rdelosv18.n69-103. Disponível em: https://ojs.revistadelos.com/ojs/index.php/delos/article/view/5912. Acesso em: 28 maio. 2026.

MARQUES, Thiago Domingos; GONÇALVES, Alexandre Leopoldo. Grafos aplicados à análise decisória em patentes: uma abordagem visual da classificação IPC - análise de redes. Revista Delos, v. 18, n. 67, e5030, 2025. Disponível em: https://doi.org/10.55905/rdelosv18.n67-045. Acesso em: 29 maio 2025.

MARQUES, Thiago Domingos; GONÇALVES, Alexandre Leopoldo. Uma revisão integrativa para sistemas de busca por patentes similares utilizando IA: avanços, desafios e aplicações. *CIKI/ICKM*. 2023. Disponível em: https://proceeding.ciki.ufsc.br/index.php/ciki/article/view/1418/826. Acesso em: 06 mar. 2025.

MARTINS, Claudia A.; FRANCISCO, Rafaela S.; FARIAS, Henrique C.. Classificação e visualização de dados de patentes. *In*: ESCOLA REGIONAL DE INFORMÁTICA DE MATO GROSSO (ERI-MT), 21. , 2021, Evento Online. Anais [...]. Porto Alegre: Sociedade Brasileira de Computação, 2021. p. 116-119. ISSN 2447-5386. DOI: https://doi.org/10.5753/eri-mt.2021.18235.

MIKOLOV, Tomas; CHEN, Kai; CORRADO, Greg; DEAN, Jeffrey. Efficient Estimation of Word Representations in Vector Space. In: *INTERNATIONAL CONFERENCE ON LEARNING REPRESENTATIONS (ICLR)*, 1., 2013, Scottsdale, Arizona, USA. *Workshop Track Proceedings.* [S.l.: s.n.], 2013.

MIAO, Ran; CHEN, Xueyu; HU, Liang; ZHANG, Zhifei; WAN, Minghua; ZHANG, Qi; ZHAO, Cairong. *PatSTEG: modeling formation dynamics of patent citation networks via the semantic-topological evolutionary graph*. In: IEEE INTERNATIONAL CONFERENCE ON DATA MINING, 2023, Shanghai. Anais [...]. Piscataway: IEEE, 2023. p. 1312–1317. DOI: https://doi.org/10.1109/ICDM58522.2023.00153. Acesso em: 5 jul. 2025.

MORESI, Eduardo Amadeu D.; PINHO, Isabel; HEDLER, Helga Cristina. Análise qualitativa de informações registradas em patentes. *Investigação Qualitativa em Educação: Avanços e Desafios*, v. 12, p. 1-10, 2022. DOI: https://doi.org/10.36367/ntqr.12.2022.e616.

MORAES, Lavínia de Carvalho et al. Análise de ambiguidade linguística em modelos de linguagem de grande escala (LLMs). Revista Texto Livre, Minas Gerais, v. 16, n. 4, p. 1–23, 20 dez. 2024. Disponível em: https://doi.org/10.1590/1983--3652.2025.53181. Acesso em: 16 jul. 2025.

Revista ARACÊ
ISSN: 2358-2472

NASCIMENTO, T.C, ROJAS CAJAVILCA, E. S., TELES SANTOS, A. Systematization of a Model of Technological Prospection With the Spacenet and Iramuteq Tools: application to the bank of green patent data of the phosphorus element. Cadernos de Prospecção, 12(3), 563-575. Universidade Federal do Oeste da Bahia - UFOB, Barreiras, BA, Brasil. 2019.

OUYANG, Xin; SUN, Zhen; XU, Xinzhen. Patent system in the digital era - Opportunities and new challenges. Journal Of Digital Economy, [S.L.], v. 1, n. 3, p. 166-179, dez. 2022. Elsevier. 2022

PAULINO, Rita de Cássia Romeiro. A interpretação de grafos como imagens complexas em tempos de pandemia de COVID-19 no Brasil. Asas da Palavra, v. 17, n. 1, p. 43-51, jan./jun. 2020.

PINTO, Adilson Luiz; SILVA, Armando Malheiro da; SENA, Priscila Machado Borges. Ontologias baseadas na visualização da informação das redes sociais. *Prisma.com*, Universidade do Porto, Portugal, 2010. Disponível em: https://www.prisma.com. Acesso em: 31 maio 2025.

PIRES, E. A.; RIBEIRO, N. M.; QUINTELLA, C. M. Sistemas de Busca de Patentes: análise comparativa entre Espacenet, Patentscope, Google Patents, Lens, Derwent Innovation Index e Orbit Intelligence. Cadernos de Prospecção, v. 13, n. 1, p. 13-29, 2020. DOI: 10.9771/cp.v13i1.35147.

REYMOND, David; DEMATRAZ, Jessica. *Using networks in patent exploration: application in patent analysis: the democratization of 3D printing*. Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação, v. 19, n. 40, p. 117-144, mai./ago., 2014

RISCH, Julian; KRESTEL, Ralf. Domain-specific word embeddings for patent classification. Data Technologies and Applications, [S.l.], v. 53, n. 1, p. 108–122, mar. 2019. DOI: 10.1108/DTA-01-2019-0002.

SANTOS, André Moraes dos; QUONIAM, Luc; KNIESS, Claudia Terezinha; REYMOND, David. *Ferramentas para extração e análise de informações em base de patentes: uma aplicação para o modelo de hélice quíntupla*. Anais do III SINGEP e II S2IS – São Paulo – SP – Brasil – 09, 10 e 11 de novembro de 2014. Disponível em: 1. Acesso em: 31 maio 2025.

SOUZA, Jackson Wilke da Cruz; SEMCOVICI, Pedro; PARDO, Thiago Alexandre Salgueiro. Proposta de algoritmo de classificação automática de papéis semânticos em português no âmbito do modelo Abstract Meaning Representation. Texto Livre, Belo Horizonte-MG, v. 18, p. e55346, 2025. DOI: 10.1590/1983-3652.2025.55346. Disponível em: https://periodicos.ufmg.br/index.php/textolivre/article/view/55346. Acesso em: 16 jul. 2025.

SOUZA, Luiz Fernando Spillere de; GONÇALVES, Alexandre Leopoldo; SOUZA, Joao Artur De. Utilização prática de Word Embedding aplicada à classificação de texto. Ciki, 2021, Florianópolis: Universidade Federal de Santa Catarina, 2021.

JIANG, Hongxun; FAN, Shaokun; ZHANG, Nan; ZHU, Bin. Deep learning for predicting patent application outcome: the fusion of text and network embeddings. *Journal of Informetrics*, v. 17, n. 2, p. 101402, 2023.

VILLA, Anna Maria; WIRZ, Manuel. A sequential patent search approach combining semantics and artificial intelligence to identify initial State-of-the-Art documents. World Patent Information, [S.L.], v. 68, p. 102096, mar. 2022. Elsevier BV. http://dx.doi.org/10.1016/j.wpi.2022.102096.

Revista ARACÊ
ISSN: 2358-2472

WITSCHARD, Daniel; JUSUFI, Ilir; MARTINS, Rafael M; KUCHER, Kostiantyn; KERREN, Andreas. Interactive optimization of embedding-based text similarity calculations. Information Visualization, [S.L.], v. 21, n. 4, p. 335-353, 3 ago. 2022. SAGE Publications. http://dx.doi.org/10.1177/14738716221114372.

WOLSKI, Luciano Zamperetti; PIZONI, Willian Aurélio; GONÇALVES, Alexandre Leopoldo. Modelo de classificação de patentes baseado em técnicas de engenharia de conhecimento. In: CONGRESSO INTERNACIONAL DE CONHECIMENTO E INOVAÇÃO – CIKI, 2022. Disponível em: https://proceeding.ciki.ufsc.br/index.php/ciki/article/view/1254/700. Acesso em: 29 maio 2025.