INSTITUTO NACIONAL DA PROPRIEDADE INDUSTRIAL

DANIEL BARROS JUNIOR

METODOLOGIA E IMPLEMENTAÇÃO DE AGRUPAMENTO DE PEDIDOS DE PATENTES EM TÓPICOS TÉCNICOS

Rio de Janeiro

2024

Daniel Barros Junior

Metodologia e implementação de agrupamento de pedidos de patentes em tópicos técnicos

Tese apresentada, como requisito parcial para obtenção do título de Doutor, ao Programa de Pós-Graduação em Propriedade Intelectual e Inovação, do Instituto Nacional da Propriedade Industrial.

Orientador: Prof. Dr. Celso Luiz Salgueiro Lage

Coorientadora: Dra. Catia Valdman

Rio de Janeiro

2024

Ficha catalográfica elaborada pela Biblioteca de Propriedade Intelectual e Inovação – INPI
Bibliotecário Responsável Técnico Evanildo Vieira dos Santos CRB7-4861

B277 Barros Junior, Daniel.
Metodologia e implementação de agrupamento de pedidos de patentes em tópicos técnicos. -- 2024.

121 f. ; figs.; tabs.

Tese (Doutorado em Propriedade Intelectual e Inovação) - Academia de Propriedade Intelectual Inovação e Desenvolvimento, Divisão de Programas de Pós-Graduação e Pesquisa, Instituto Nacional da Propriedade Industrial - INPI, Rio de Janeiro, 2024.

Orientador: Prof. Dr. Celso Luiz Salgueiro Lage.
Coorientadora: Dra. Catia Valdman.

1. Propriedade industrial – Patente. 2. Patente – Exame – Similaridade semântica. 3. Patente – Exame – Modelos de linguagens. 4. Redes Neurais. 5. BERTopic. I. Instituto Nacional da Propriedade Industrial (Brasil).

CDU: 347.771:001.4

14/05/2025
Data

DANIEL BARROS JUNIOR:56639503034
Digitally signed by DANIEL BARROS JUNIOR:56639503034
Date: 2025.05.14 22:52:10 -03'00'
Assinatura

Daniel Barros Junior

Metodologia e implementação de agrupamento de pedidos de patentes em tópicos técnicos

Tese apresentada, como requisito parcial para obtenção do título de Doutor, ao Programa de Pós-Graduação em Propriedade Intelectual e Inovação, do Instituto Nacional da Propriedade Industrial.

Aprovada em 16 de dezembro de 2024.

Banca Examinadora:

Prof. Dr. Celso Luiz Salgueiro Lage (orientador)
Instituto Nacional da Propriedade Industrial

Dra. Catia Valdman (coorientadora)
Instituto Nacional da Propriedade Industrial

Prof. Dr. Alexandre Guimarães Vasconcellos
Instituto Nacional da Propriedade Industrial

Profa. Dra. Genizia Islabão de Islabão
Instituto Nacional da Propriedade Industrial

Prof. Dr. Erik Schüler
Instituto Federal de Educação, Ciência de Tecnologia do Rio Grande do Sul

Dr. Vagner Luís Latsch
Instituto Nacional da Propriedade Industrial

Dr. José Aguiar Coelho Neto
Instituto Nacional da Propriedade Industrial

Rio de Janeiro

2024

DEDICATÓRIA

Dedico esta tese a minha esposa Rosângela e meu filho André sem os quais nada disso seria possível.

AGRADECIMENTOS

Agradeço a minha família em especial minha esposa Rosângela, meu filho André, minha mãe Maria Julia (*in memoriam*), meu pai Daniel (*in memoriam*) e meu irmão Sérgio. Todos sempre me apoiaram e auxiliaram.

Agradeço aos compadres, aos afiliados e aos amigos que sempre ajudaram em todas as horas.

Agradeço ao meu orientador Prof. Celso e a minha coorientadora Catia, por toda a paciência e dedicação.

Agradeço a Academia do INPI que possibilitou a realização desta tese a distância, disponibilizando um curso de alta qualidade para todas as pessoas do Brasil.

Daniel Barros Junior

“O que sabemos é uma gota; o que ignoramos é um oceano. Mas o que seria o oceano se não infinitas gotas?”

Isaac Newton

RESUMO

BARROS JUNIOR, Daniel. **Metodologia e implementação de agrupamento de pedidos de patentes em tópicos técnicos**. 2024. 121 f. Tese (Doutorado em Propriedade Intelectual e Inovação) – Instituto Nacional da Propriedade Industrial, Rio de Janeiro, 2024.

A presente tese propõe uma metodologia para agrupar pedidos de patentes em tópicos técnicos específicos utilizando informações sobre os pedidos e seus conteúdos técnicos, tais como: reivindicações, relatório descritivo e resumo. A principal motivação para o agrupamento dos pedidos em tópicos técnicos é o fornecimento de subsídios aos chefes das divisões técnicas de patentes para auxiliar na distribuição de pedidos similares entre os examinadores de patentes. Foram utilizados modelos de transformadores de frases buscando suas similaridades semânticas. Para isso, foram exploradas as ferramentas disponibilizadas pela técnica de modelagem de tópicos BERTopic, a qual possui uma grande flexibilidade e pode ser configurada para utilizar módulos desenvolvidos na análise de textos. A validação dos resultados é baseada na obtenção de pedidos de patentes em conjunto com pedidos de suas famílias, pedidos citados e pedidos similares. Observou-se o desempenho dos modelos de linguagem em agrupar pedidos de uma determinada área do conhecimento, juntamente com os pedidos de suas famílias, citados e similares. O modelo de linguagem selecionado foi o que obteve o melhor desempenho com relação a proximidade dos pedidos do conjunto inicialmente selecionado (ATInicial) e as suas famílias nas três situações, ao final do treinamento, com os dados de treinamento e com os dados de teste. O nível de agrupamento dos pedidos do conjunto ATInicial com pedidos das suas famílias foi utilizado como balizador para indicar a eficiência do modelo de linguagem no agrupamento de pedidos semelhantes.

Palavras-chave: similaridade semântica, patentes, redes neurais, BERTopic

ABSTRACT

BARROS JUNIOR, Daniel. **Methodology and implementation of grouping patent applications into technical topics**. 2024. 121 f. Tese (PhD in Intellectual Property and Innovation) – Instituto Nacional da Propriedade Industrial, Rio de Janeiro, 2024.

The present thesis proposes a methodology to group patent applications into specific technical topics using information about the applications and their technical contents, such as: claims, descriptive report and summary. The main motivation for grouping applications into technical topics is to provide subsidies to the heads of the patent technical divisions to assist in the distribution of similar applications among patent examiners. Sentence transformer models were used, seeking their semantic similarities. For this, the tools made available by the BERTopic topic modeling technique were explored, which has great flexibility and can be configured to use modules developed in text analysis. Validation of the results is based on obtaining patent applications in conjunction with applications from their families, cited applications, and similar applications. The performance of the language models in grouping requests from a certain area of knowledge was observed, together with the requests of their families, cited and similar. The selected language model was the one that obtained the best performance in relation to the proximity of the requests of the initially selected set (ATInicial) and their families in the three situations, at the end of the training, with the training data and with the test data. The level of grouping of the requests of the ATInicial set with requests from their families was used as a marker to indicate the efficiency of the language model in grouping similar requests.

Keywords: semantic similarity, patents, neural networks, BERTopic

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

INTRODUÇÃO

Uma patente é um direito conferido pelo Estado e dá ao seu titular exclusividade de exploração de uma tecnologia (BARBOSA, 2010). Para que seja obtido este direito, o exame da patente precisa ser realizado pelo Instituto Nacional da Propriedade Industrial (INPI). O atraso no exame é um dos motivos que gera a demora da decisão das patentes e prejudica a sociedade de modo geral. Portanto, alternativas que auxiliem na otimização do exame de patentes são relevantes.

O trâmite de um pedido de patentes apresenta diversas etapas que podem variar conforme o caso. A Figura 1 apresenta uma descrição simplificada de algumas destas etapas:

- Depósito – no momento do depósito o pedido recebe um número e entra na fila para que seja realizado o exame formal. Este exame se preocupa com aspectos legais e de formato dos documentos, sem análise de conteúdo técnico;
- Classificação – passados os prazos legais de sigilo, o pedido será publicado disponibilizando seus conteúdos. Antes da publicação, ele precisa ser classificado por um examinador de patentes;
- Administrativo – após a publicação, passados os prazos e eventos legais desta etapa, o pedido entra para a fila de exame. Durante o período em que o pedido aguarda o exame, a equipe administrativa controla se todos os pagamentos e as formalidades exigidas estão em dia;
- Distribuição para exame – observados os critérios de ordenação das filas de exame, os pedidos são distribuídos pelos chefes de divisão. Cada divisão é responsável por um conjunto de classificações e cada pedido recebe sua classificação no início do processo. **A presente tese busca fornecer subsídios aos chefes para a distribuição dos pedidos para os examinadores, agrupando pedidos que apresentem similaridades**;
- Exame – o pedido é analisado e é emitido um primeiro parecer, que pode ser já uma decisão ou uma avaliação que precisa ser respondida pelo requerente.

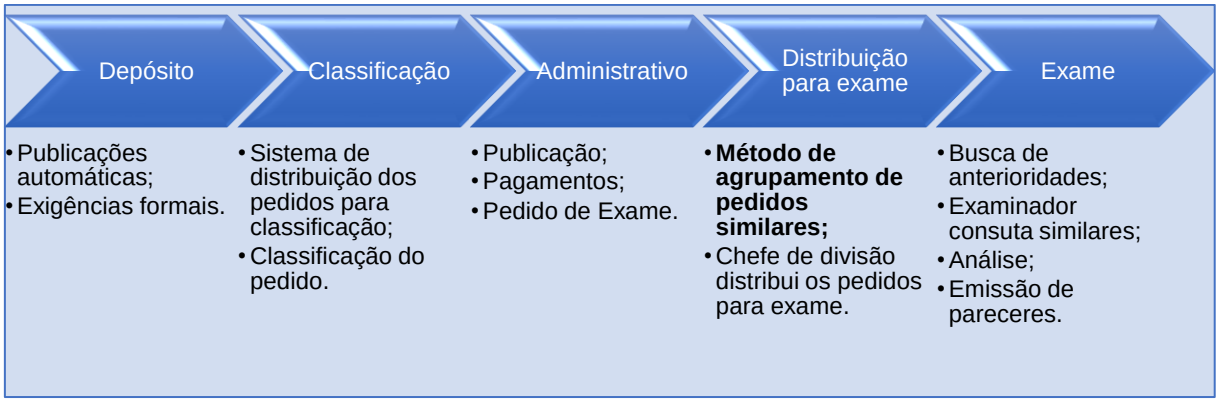


Figura 1 – Fluxo simplificado de exame
Fonte: Próprio autor

O atraso no exame dos pedidos de patentes é ocasionado por diversos fatores, como o grande volume de pedido de patentes depositados, o número de examinadores ser menor do que o necessário para atender a demanda, os trâmites de exame de patentes poderem apresentar várias etapas de pareceres e petições de resposta, entre outros motivos. Portanto, a otimização de qualquer etapa do exame dos pedidos de patentes é sempre relevante.

Em uma recente tese de doutorado defendida na Academia do INPI, o autor construiu um sistema que busca determinar parâmetros para atribuir aos pedidos níveis de dificuldades para o exame, auxiliando aos chefes de divisão na distribuição dos pedidos para os examinadores de forma mais equilibrada (MOREIRA JR., 2021). São exemplos como este que servem de inspiração para que sejam realizados novos trabalhos que busquem a otimização do exame de pedidos de patentes.

Neste sentido, observou-se que atualmente os pedidos chegam às divisões técnicas somente pela sua classificação, cabendo aos chefes de divisão separar os pedidos para os examinadores e eventualmente encaminhar para outras áreas quando as classificações não são da sua área de trabalho. Uma vez identificados os documentos de patentes a serem avaliados na sua área, os chefes devem então distribuir estes pedidos para exame, muitas vezes de modo aleatório. Uma forma de minimizar este problema seria agrupar os pedidos que apresentam similaridades nos seus conteúdos técnicos e seus metadados.

O termo metadados vem cada vez mais sendo utilizado para definir a informação sobre os dados, tendo como uma de suas finalidades o auxílio na localização e organização dos dados propriamente ditos (GOOD, 2003).

A presente tese pesquisa formas de agrupar os pedidos de patentes principalmente pelos seus conteúdos técnicos comparando os pedidos uns com os outros.

Um ponto de vital importância é determinar o conjunto de pedidos que deve ser comparado, visando o equilíbrio entre a assertividade e o tempo de execução para realizar esse agrupamento de pedidos similares.

Utilizam-se dados de patentes dos Estados Unidos da América (US) para a geração e teste dos modelos. A ideia é criar uma metodologia que, no futuro, possa ser adaptada para a realidade do INPI, colaborando no aumento de subsídios para que os chefes de divisão possam distribuir os pedidos para os examinadores.

O objetivo principal da tese é propor uma metodologia para construir modelos que agrupem pedidos em tópicos técnicos específicos, utilizando informações sobre os pedidos e seus conteúdos técnicos, tais como reivindicações, relatório descritivo e resumo.

Dados de patentes US foram utilizados nas etapas de treinamento e de teste para validar os modelos. Essa abordagem foi adotada devido à grande quantidade de documentos US disponíveis. A validação da metodologia de construção de um modelo de linguagem para uma determinada área justifica a aplicação da metodologia para outras áreas e o desenvolvimento de estudos para aprimorar a pesquisa.

Para chegar neste objetivo, foi preciso definir e realizar outras etapas englobadas nos seguintes objetivos específicos: a) identificar metadados e dados que se mostrem relevantes para agrupar os pedidos e validar os modelos; b) determinar condições de contorno que podem ser utilizadas para viabilizar a execução dos algoritmos de treinamento e teste dos modelos; c) realizar testes utilizando conjuntos de dados selecionados para verificar a efetividade dos modelos; d) construir e testar um modelo de tópicos técnicos para agrupar documentos de uma determinada área.

Os modelos desenvolvidos com a metodologia descrita agrupam os pedidos de patentes similares. Entende-se ser impossível para o examinador lembrar dos detalhes de cada pedido, bem como de quais pedidos já examinados poderiam

contribuir para realizar o exame do pedido atual. Portanto, saber quais pedidos um examinador já examinou e quais são similares, além de dar agilidade ao exame, auxilia na busca em manter a coerência das decisões entre os pedidos, proporcionando maior segurança jurídica às decisões.

Em outras palavras, estes modelos auxiliarão em atividades realizadas durante o exame dos pedidos. Por exemplo:

- Reunir pedidos similares que foram examinados por outros examinadores, possibilitando o acesso a pontos de vista diferentes;
- Detectar pedidos similares ao pedido que está sendo examinado;
- Catalogar pedidos similares que já foram examinados e podem ser utilizados para consulta de decisões e aproveitamentos de buscas;
- Reservar pedidos considerados similares para exames futuros, que ainda não possam ser examinados por algum motivo administrativo/processual - por exemplo, terem a data de depósito mais recente;
- Evitar que um mesmo examinador tome decisões muito diferentes para pedidos que apresentam similaridade ou que, pelo menos, o examinador tenha em mente o que deve ser argumentado para mostrar as diferenças de opinião entre os pedidos, haja vista serem similares, mas não iguais.

Os modelos baseados em tópicos agrupam documentos que são representados por características que os tornam similares. Sendo assim, neste trabalho o critério para considerar os pedidos como similares foi o agrupamento em um mesmo tópico técnico ou em tópicos técnicos adjacentes.

Objetivos

Objetivo geral da tese:

Propor uma metodologia para construir modelos de linguagem que agrupem pedidos em tópicos técnicos específicos, utilizando os conteúdos técnicos dos pedidos de patentes, tais como: reivindicações, relatório descritivo e resumo, buscando gerar subsídios para a distribuição dos pedidos para exame.

Objetivos específicos:

I – Identificar metadados e dados que se mostrem relevantes para agrupar os pedidos e validar os modelos;

II – Determinar condições de contorno que podem ser utilizadas para viabilizar a execução dos algoritmos de treinamento e teste dos modelos;

III – Realizar testes utilizando conjuntos de dados selecionados para verificar a efetividade dos modelos;

IV – Construir e testar um modelo de tópicos técnicos para agrupar documentos de uma determinada área.

Presume-se que a metodologia proposta seja capaz de permitir a geração de modelos para diversas áreas do conhecimento. Considera-se que os modelos são dependentes das palavras e cada área do conhecimento possui um conjunto de palavras que irá diferenciá-la de outra área.

Entende-se que o retreinamento dos modelos só será necessário quando for detectada uma queda de desempenho ou ocorrerem modificações significativas na área específica para a qual o modelo foi treinado. Um exemplo de modificação significativa seria uma reestruturação no sistema de classificação da área para a qual o modelo foi treinado, com a inserção de novos conceitos e, por consequência, um novo conjunto de palavras para os documentos que identificam a área.

Motivação da tese

Os conteúdos técnicos dos pedidos de patentes apresentam ao longo do texto diversos conceitos e é necessário identificar quais palavras definem o tópico mais relevante. Muitas vezes, um examinador busca por palavras-chave que definam o conteúdo principal que trata o pedido e utiliza estas palavras para encontrar pedidos similares.

Questão de pesquisa

Um ponto relevante para este trabalho é auxiliar no mapeamento de pedidos que possam ser examinados em conjunto ou que sirvam como subsídio ao exame de um determinado pedido de patente. Identificar e mapear pedidos de patentes similares

quando é realizado manualmente, reduz a escala de documentos que podem ser encontrados.

Importante esclarecer que foram considerados como metadados todas as informações que estão disponíveis que não sejam de conteúdo técnico dos pedidos. Como exemplo de metadados pode-se citar: requerentes, titulares, inventores, classificações, idiomas, prioridades e datas. Por outro lado, os dados são basicamente o conteúdo técnico dos pedidos, contidos no relatório descritivo, nas reivindicações e no resumo.

A Figura 2 apresenta dados e metadados que podem ser utilizados na metodologia desta tese. Atualmente, a quantidade de informações relacionadas aos pedidos de patentes está cada vez mais diversificada e acessível. Os próprios textos presentes na estrutura dos símbolos dos sistemas de classificação (por exemplo, IPC e CPC, ver apêndice A) são fontes de informações sobre os pedidos de patentes. Uma abordagem para o treinamento de modelos poderia ser utilizar as palavras que compõem os textos do próprio esquema de classificação. O examinador ao classificar um pedido procura palavras que estão no pedido e no esquema de classificação para influenciar a sua decisão.

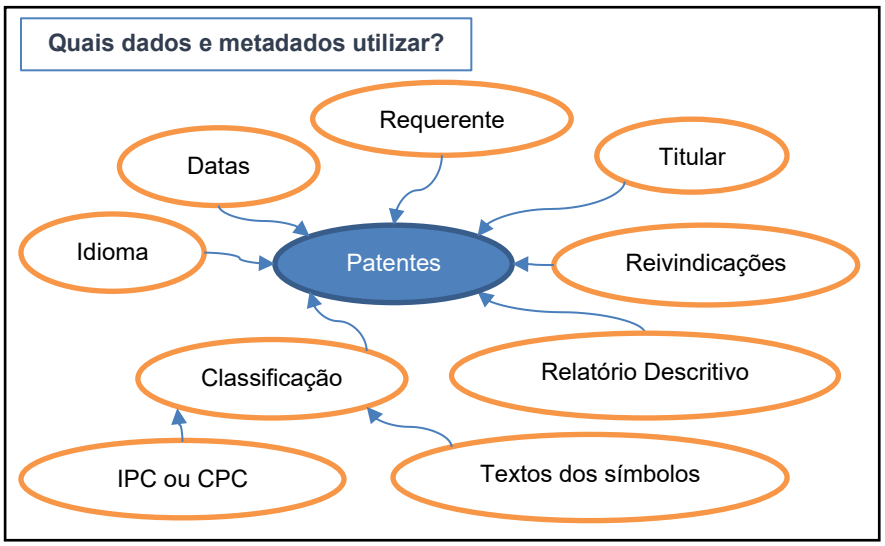


Figura 2 – Exemplos de dados e metadados que podem ser utilizados
Fonte: Próprio autor

Foram escolhidos apenas os textos das reivindicações para a elaboração do algoritmo apresentado na metodologia desta tese. Esta premissa foi baseada no fato de que nas reivindicações tem-se todas as palavras e/ou frases que definem o assunto que é objeto de proteção do pedido. Segundo Hain *et al.* (2022), o uso do resumo seria mais adequado para encontrar tecnologias semelhantes e o uso das reivindicações seria mais adequado para encontrar infrações entre patentes. Isso contribui com a decisão de utilizar as reivindicações para o presente trabalho, pois o objetivo aqui é auxiliar na realização do exame. Entende-se que são as reivindicações que irão determinar a proximidade entre as patentes e as limitações dos objetos a serem explorados por seus detentores.

Optou-se pelo uso de documentos em inglês devido à grande quantidade de pedidos. Mesmo os documentos brasileiros são encontrados com uma tradução do português para o inglês; ainda que a tradução tenha sido por máquina, os textos disponíveis possuem coerência. Considerando que as técnicas usadas são baseadas em redes neurais, um grande volume de documentos foi necessário para a realização de treinamentos e de teste dos modelos.

A Figura 3 cita exemplos de abordagens que auxiliam na solução do problema. Nesta tese, para o agrupamento de pedidos foram utilizadas técnicas baseadas em redes neurais, mas outros procedimentos poderiam ser utilizados para aumentar o desempenho do processo.

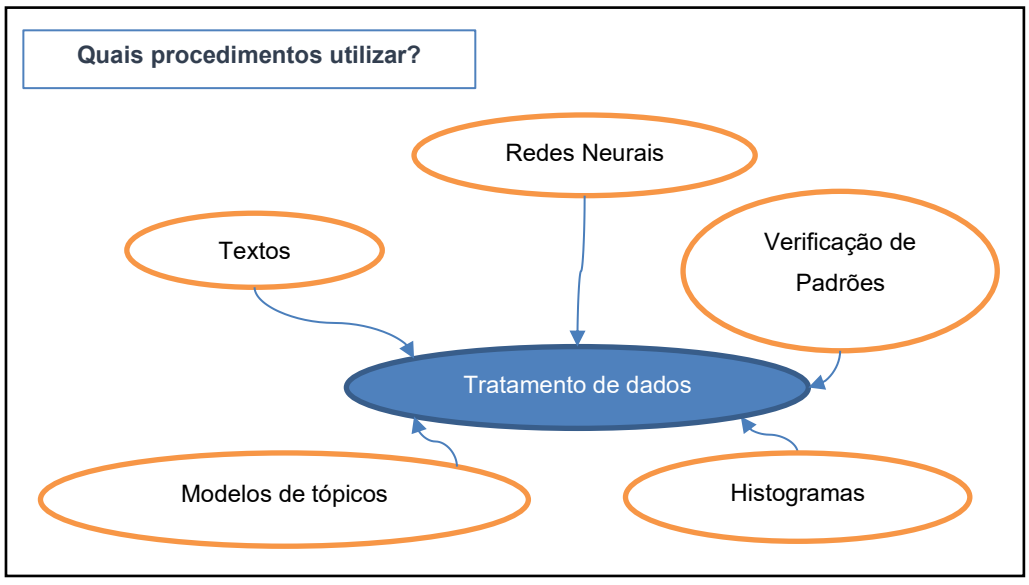


Figura 3 – Exemplos de procedimentos que podem ser utilizadas
Fonte: Próprio autor

Na Figura 4 estão algumas alternativas que podem ser utilizadas para a verificação dos resultados, a fim de que um modelo possa ser validado para outras áreas. A correta interpretação dos resultados é essencial.

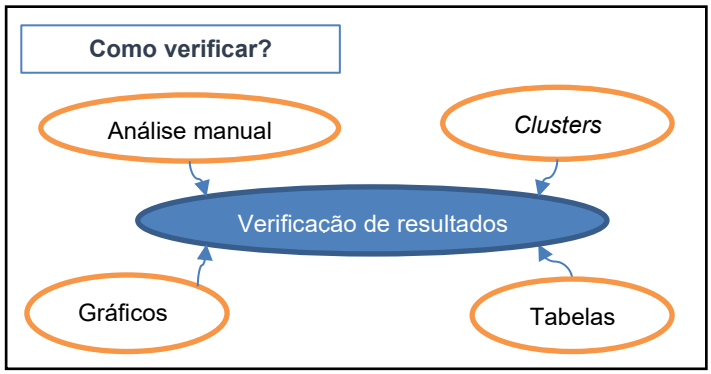


Figura 4 – Exemplos de opções para verificação dos resultados
Fonte: Próprio autor

Justificativa

Os sistemas de classificação de pedidos de patentes têm como objetivo principal categorizar e organizar os pedidos. Mesmo os pedidos que têm classificação semelhante apresentam palavras diferentes em seus conteúdos.

Considerando a utilização das palavras para determinar a proximidade dos pedidos, foi necessário utilizar uma grande quantidade de pedidos para que os conjuntos de treinamento dos modelos contenham uma boa diversidade de palavras utilizadas em uma determinada área.

Pedidos que tratam de assuntos semelhantes tendem a ser examinados pelo mesmo grupo de examinadores. A principal motivação para a realização deste trabalho está em apresentar uma metodologia que torne viável encontrar pedidos de patentes que apresentam similaridades ao pedido de patente que está pronto para ser analisado.

Atualmente, ao inserir um pedido na carga de trabalho de um examinador, o sistema interno de controle de produção do INPI sugere pedidos similares. Este sistema baseia-se em uma heurística de comparação que leva em consideração metadados tais como: requerente, inventores, classificações e prioridades. O sistema procura

apresentar aos chefes de divisão quais pedidos possuem relação de forma mais direta com o pedido inserido na carga do examinador, no intuito de sugerir que estes pedidos similares sejam também inseridos para o mesmo examinador.

A busca dessa nova abordagem é criar uma relação entre pedidos do mesmo requerente ou de diferentes requerentes que podem apresentar uma relação entre o objeto reivindicado, indicando essa relação aos chefes de divisão. Estes chefes podem utilizar isso como uma ferramenta para agilizar a distribuição dos pedidos.

Do ponto de vista dos examinadores, a visualização desta relação entre os pedidos pode gerar inúmeros benefícios. O examinador pode vir a utilizar as anterioridades já encontradas nos pedidos similares. O examinador também pode se concentrar em uma área mais específica, visto que ele começa a receber pedidos que são próximos uns dos outros. Com isso, a sua produtividade tende a aumentar, pois cada vez mais ele se torna um especialista. Procurou-se apresentar estes benefícios na Tabela 1.

Tabela 1 – Benefícios

Ator	Benefícios
Chefes de Divisão	Auxílio na distribuição dos pedidos
Examinador	Concentração em uma determinada área
	Aproveitamento de anterioridades
	Agilidade no exame
	Aumento de produtividade
INPI/DIRPA	Exames realizados mais rapidamente
	Aumento na segurança jurídica, considerando que possivelmente ocorrerão decisões mais semelhantes para pedidos similares

Fonte: Próprio autor

Na literatura, existem trabalhos que procuram fazer o mapeamento das áreas de conhecimento contidas em pedidos de patentes e demonstrar a proximidade entre estes pedidos, por exemplo: AHARONSON; SCHILLING, 2016; MASTROGIORGIO; GILSING, 2016. Outros trabalhos utilizam técnicas para seleção e organização de documentos, por exemplo: BENNER; WALDFOGEL, 2008; BOON; PARK, 2005; COOKE; GILLAM, 2017; ECKARDT *et al.*, 2010, 2020; GERKEN; MOEHRLE, 2012;

GUAN; YAN, 2016; HSU *et al.*, 2021; JEONG; LEE, 2015; LEE; HAN; SOHN, 2015; WALSH; LEE; JUNG, 2016; YOON; KIM, 2012; YOON; PARK; KIM, 2013.

Entretanto, os trabalhos encontrados estão focados principalmente em resolver problemas da indústria, tentando mapear as tecnologias exploradas pelos fabricantes. Tal enfoque difere do objetivo principal da presente proposta, o qual é mapear e correlacionar pedidos de patentes que tenham conteúdos semelhantes, permitindo que sejam examinados em um mesmo momento.

Delimitações do tema

A presente tese visa validar a metodologia do uso de modelos de linguagem para agrupar pedidos de patentes similares. Para que o uso desta metodologia seja avaliado de forma consistente se faz necessário um projeto com situações que contemplem pedidos reais e a implementação em várias áreas do conhecimento. Entretanto, devido a complexidade, a aplicação da metodologia em uma escala mais ampla não está no escopo deste trabalho.

Organização da tese

A tese é composta por cinco seções, quais sejam: introdução, três capítulos e conclusão. A introdução apresenta um pequeno histórico das motivações e justificativas para a realização deste trabalho. O Capítulo 1 busca organizar e delimitar diversos conceitos que foram necessários para o melhor entendimento dos problemas e desenvolvimento de soluções. O Capítulo 2 define as características da metodologia desenvolvida. O Capítulo 3 apresenta um exemplo detalhado da implementação da metodologia proposta e outros exemplos com modificações, visando mensurar a melhora de desempenho. Ao final, são apresentadas as conclusões, baseadas nas análises dos resultados dos testes realizados.

1 FUNDAMENTAÇÃO TEÓRICA

Neste capítulo, serão apresentados conceitos básicos de propriedade intelectual detalhando os pedidos de patentes, bases de patentes onde podem ser adquiridos os documentos de patentes, uma abordagem superficial de inteligência artificial (IA), conceitos de similaridade semântica entre documentos, conceitos de aprendizado de máquina, conceitos de processamento de linguagem natural e conceitos de técnicas de modelagem de tópicos essenciais para metodologia proposta na tese.

1.1 PROPRIEDADE INTELECTUAL

Segundo o Prof. Denis Barbosa, a propriedade intelectual abrange diversos ativos intelectuais, entre eles estão as invenções em todos os domínios da atividade humana (BARBOSA, 2010, p. 10). Englobado neste tipo de proteção intelectual estão as patentes, as quais são o foco principal deste trabalho e são consideradas um direito que dá ao titular a exclusividade da exploração de uma determinada tecnologia (BARBOSA, 2010, p. 295).

A Figura 5 apresenta de forma simplificada os ativos intangíveis que constituem a propriedade intelectual, a qual é subdividida em propriedade industrial, direito autoral e proteção *sui generis*[1]. As patentes são consideradas ativos financeiros pertencentes ao campo da propriedade industrial.

[1] Proteção sui generis são proteções de ativos intelectuais que não se enquadram em Propriedade Industrial ou Direito Autoral. Por exemplo, topografia de circuitos integrados, novas variedades de plantas, patrimônio genético e conhecimento tradicional associado.

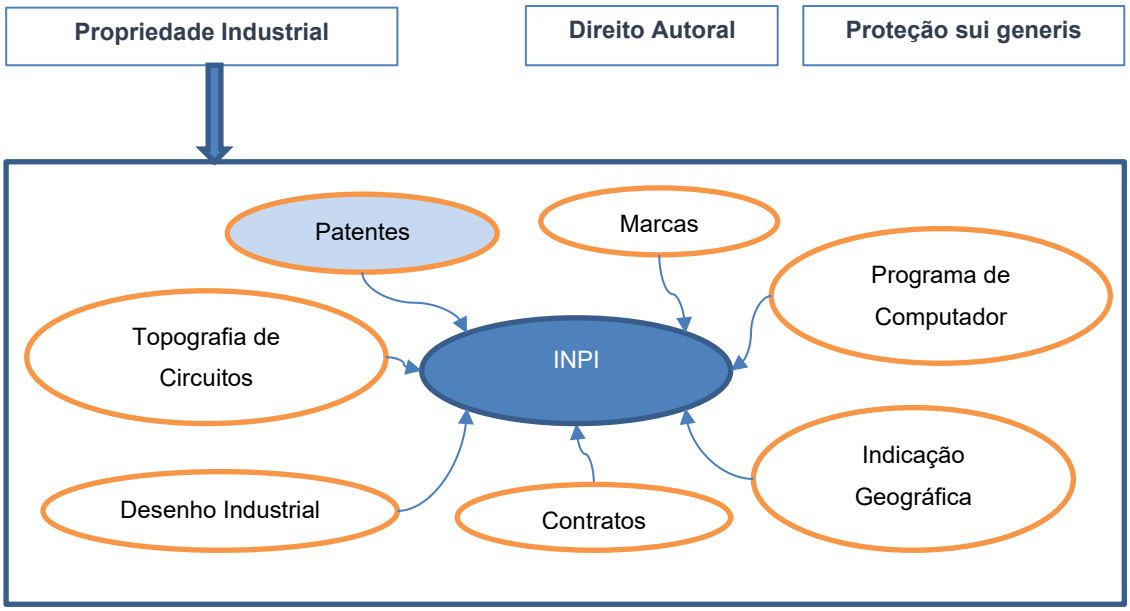


Figura 5 – Ativos Intangíveis da Propriedade Intelectual
Fonte: INPI

1.1.1 Pedido de patente

O processo de pedidos de patentes é composto por diversas etapas regidas pela Lei de Propriedade Industrial (LPI) (BRASIL, 1996), que define os conteúdos que precisam ser apresentados e os mais diversos procedimentos, conforme já apresentado no fluxo simplificado na Figura 1.

Os conteúdos técnicos do pedido de patente estão no relatório descritivo, nas reivindicações, nos desenhos e no resumo. Informações bibliográficas podem ser relacionadas ao pedido, para que ele seja corretamente catalogado e mais facilmente localizado.

1.1.2 Família de patentes

Uma família de patentes é um conjunto de pedidos de patentes depositados e referentes a uma mesma invenção (INPI, 2021). As bases de dados apresentam conceitos diferenciados de família.

O DOCDB é a base de dados de documentação do *European Patent Office* (EPO) e contém dados bibliográficos, citações, resumos e a família de patentes simples. O conceito de família de patentes simples consiste em ter uma coleção de documentos

de patentes que são considerados como cobrindo uma única invenção. O conteúdo técnico coberto pelos pedidos é considerado idêntico. Os membros de uma família de patentes simples terão exatamente as mesmas prioridades[2].

O INPADOC é uma base de dados do EPO que apresenta um conceito de família de patentes estendida, o qual corresponde a uma coleção de documentos de patentes que abrangem uma tecnologia. O conteúdo técnico abrangido pelos pedidos de patentes é semelhante, mas não necessariamente o mesmo. Os membros de uma família de patentes estendida terão pelo menos uma prioridade em comum[3].

Para a metodologia proposta, o conceito de família utilizado pode ser qualquer um deles. Mas o conceito de família estendida disponível no INPADOC parece ser mais adequado, pois baseia-se na complementaridade entre os documentos e não na igualdade dos conteúdos.

1.1.3 Estado da técnica

O estado da técnica é constituído por tudo aquilo tornado acessível ao público antes da data de depósito do pedido de patente (INPI, 2021).

Todos os documentos que se tornaram públicos antes da data do depósito do pedido podem ser considerados como estado da técnica e por consequência podem ser úteis para determinar se a patente atende aos critérios de patenteabilidade.

1.1.4 Documentos citados

Os documentos citados pelos requerentes nos pedidos de patentes devem ser utilizados para definir o estado da técnica, ao qual os pedidos de patentes pertencem.

Os documentos citados pelos especialistas que analisam os pedidos de patentes são documentos de anterioridades utilizados para demonstrar que o pedido de patentes em análise precisa ser modificado ou mesmo que este não pode ser concedido.

---

[2] Disponível em: https://www.epo.org/en/searching-for-patents/helpful-resources/first-time-here/patent-families/docdb Acesso em: 4 out, 2024.

[3] Disponível em: https://www.epo.org/en/searching-for-patents/helpful-resources/first-time-here/patent-families/inpadoc Acesso em: 4 out, 2024.

Entre os documentos citados estão os subsídios relacionados aos pedidos por qualquer interessado no pedido e que devem ser analisados durante o processo de exame técnico do pedido de patentes.

1.1.5 Dados e metadados dos pedidos de patentes

Segundo Good, (2003) o termo metadados vem cada vez mais sendo utilizado para definir os “dados sobre os dados”. Todas as informações catalográficas sobre os documentos são consideradas metadados dos documentos. Por exemplo, a citação de um livro é um tipo de metadado sobre o livro.

Os metadados se caracterizam por descrever alguma informação sobre o documento sem ser os dados que constituem o documento. Em um livro os metadados são informações como idioma, autor, editora, data de edição, etc. Por outro lado, os dados de um livro são os textos de seus capítulos.

Os metadados podem ser tornados públicos com maior facilidade do que os dados dos documentos. Dados bibliográficos, que permitem ao leitor de um determinado trabalho encontrar as referências citadas, também são metadados, como por exemplo, nome do autor, universidade, empresa, país, revista de publicação (GOOD, 2003; KNIGHT, 2023).

Com relação a patentes, a data de depósito (metadado) de um pedido pode ser tornada pública muito mais facilmente do que o relatório descritivo (dado) do pedido haja vista os aspectos legais de sigilo, que são exigidos no sistema patentário.

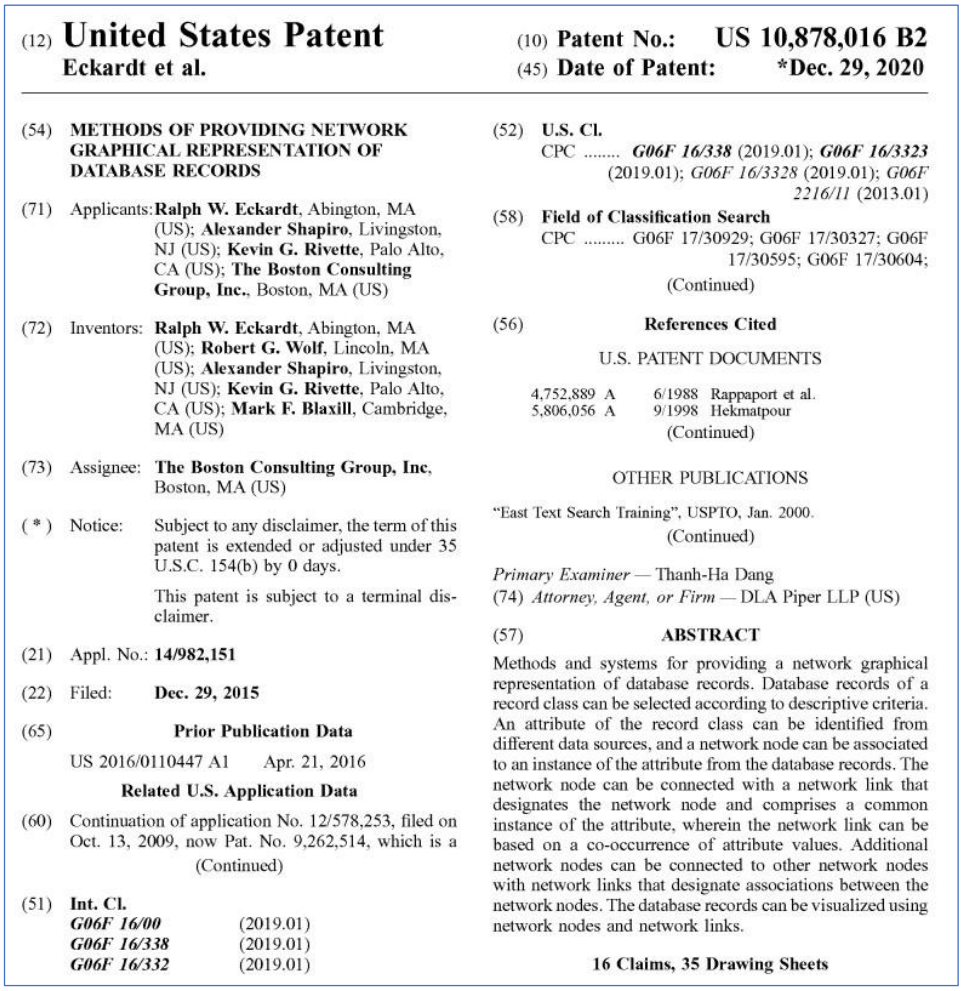
Existem padrões recomendados pela *World Intellectual Property Organization* (WIPO)[4], em português Organização Mundial da Propriedade Intelectual (OMPI), os quais buscam definir os dados bibliográficos. O padrão determinado pela OMPI, usado pelos países membros é chamado de *Internationally agreed Numbers for the Identification of (bibliographic) Data* (INID) (WIPO, 2013).

A Figura 6 apresenta uma patente concedida nos Estados Unidos da América. Na figura é possível identificar os INIDs entre parênteses ao lado de cada informação bibliográfica. O INID (10) indica o número de publicação do pedido: US10,878,016B2.

[4] Disponível em: https://www.wipo.int/ Acesso em: 4 out, 2024.

O INID (54) indica o título do pedido, no qual se pode observar que este pedido trata de métodos para fornecer representação gráfica de rede de registros de banco de dados.

Outros INIDs podem ser observados na Figura 6. Entretanto, é importante observar que alguns destes INIDs são utilizados para identificar metadados e outros para identificar dados do pedido. Por exemplo, os INIDs (54) e (57) são dados que revelam informações técnicas sobre o conteúdo, sendo respectivamente o título e o resumo do pedido.



(12) **United States Patent**
Eckardt et al.

(10) **Patent No.: US 10,878,016 B2**
(45) **Date of Patent: *Dec. 29, 2020**

(54) **METHODS OF PROVIDING NETWORK GRAPHICAL REPRESENTATION OF DATABASE RECORDS**

(71) Applicants: **Ralph W. Eckardt**, Abington, MA (US); **Alexander Shapiro**, Livingston, NJ (US); **Kevin G. Rivette**, Palo Alto, CA (US); **The Boston Consulting Group, Inc.**, Boston, MA (US)

(72) Inventors: **Ralph W. Eckardt**, Abington, MA (US); **Robert G. Wolf**, Lincoln, MA (US); **Alexander Shapiro**, Livingston, NJ (US); **Kevin G. Rivette**, Palo Alto, CA (US); **Mark F. Blaxill**, Cambridge, MA (US)

(73) Assignee: **The Boston Consulting Group, Inc**, Boston, MA (US)

(*) Notice: Subject to any disclaimer, the term of this patent is extended or adjusted under 35 U.S.C. 154(b) by 0 days.

This patent is subject to a terminal disclaimer.

(21) Appl. No.: **14/982,151**

(22) Filed: **Dec. 29, 2015**

(65) **Prior Publication Data**

US 2016/0110447 A1 Apr. 21, 2016

Related U.S. Application Data

(60) Continuation of application No. 12/578,253, filed on Oct. 13, 2009, now Pat. No. 9,262,514, which is a (Continued)

(51) **Int. Cl.**
G06F 16/00 (2019.01)
G06F 16/338 (2019.01)
G06F 16/332 (2019.01)

(52) **U.S. Cl.**
CPC ***G06F 16/338*** (2019.01); ***G06F 16/3323*** (2019.01); *G06F 16/3328* (2019.01); *G06F 2216/11* (2013.01)

(58) **Field of Classification Search**
CPC G06F 17/30929; G06F 17/30327; G06F 17/30595; G06F 17/30604;
(Continued)

(56) **References Cited**

U.S. PATENT DOCUMENTS

4,752,889 A 6/1988 Rappaport et al.
5,806,056 A 9/1998 Hekmatpour
(Continued)

OTHER PUBLICATIONS

"East Text Search Training", USPTO, Jan. 2000.
(Continued)

Primary Examiner — Thanh-Ha Dang
(74) *Attorney, Agent, or Firm* — DLA Piper LLP (US)

(57) **ABSTRACT**

Methods and systems for providing a network graphical representation of database records. Database records of a record class can be selected according to descriptive criteria. An attribute of the record class can be identified from different data sources, and a network node can be associated to an instance of the attribute from the database records. The network node can be connected with a network link that designates the network node and comprises a common instance of the attribute, wherein the network link can be based on a co-occurrence of attribute values. Additional network nodes can be connected to other network nodes with network links that designate associations between the network nodes. The database records can be visualized using network nodes and network links.

16 Claims, 35 Drawing Sheets

Figura 6 – Dados bibliográficos de um Pedido de Patente
Fonte: Site Google Patents[5]

[5] Disponível em: https://patents.google.com/ Acesso em: 4 out, 2024.

Os INIDs (51) e (52) fazem referências as classificações que foram atribuídas ao pedido, considerando o âmbito nacional e internacional dos pedidos de patentes. A classificação dos pedidos é o que permite fazer um primeiro agrupamento entre pedidos semelhantes.

1.1.6 Textos técnicos dos pedidos de patentes

Os textos técnicos do pedido de patentes são constituídos pelo título, relatório descritivo, resumo e reivindicações (WIPO, 2022).

O título se propõe a facilitar o conhecimento da área técnica em que o pedido está situado. É o título que, em um primeiro momento, chama a atenção de eventuais interessados (BARBOSA, 2010, p. 379).

O relatório descritivo precisa descrever de forma clara e suficiente o objeto que será reivindicado, tornando possível a sua realização por uma pessoa que possua conhecimento técnico na área (BARBOSA, 2010; BRASIL, 1996).

O resumo deve apresentar os detalhes que melhor definem o objeto reivindicado (WIPO, 2022). Algumas bases de dados se diferenciam por possuírem resumos que são refeitos por especialistas da área em questão. São bases de dados que mantém os resumos originais, mas criam seus próprios resumos, os quais definem de forma mais detalhada as características do pedido de patente.

As reivindicações precisam ser fundamentadas no relatório descritivo e são elas que traçam o escopo jurídico da exclusividade a que se propõe o pedido de patente perante a lei. É nas reivindicações que o inventor da patente deve apresentar as características particulares do pedido, sempre de modo claro e preciso (BARBOSA, 2010; BRASIL, 1996).

1.1.7 Sistema de classificação

Existem diversos sistemas de classificação, inclusive cada escritório pode criar o seu conjunto de classificações. Entretanto, este trabalho se limitou a apresentar a

Classificação Internacional de Patentes[6] (CIP), sendo que, frequentemente, é utilizado o termo em inglês *International Patent Classification*[7] (IPC). O principal objetivo da classificação é proporcionar uma ferramenta de busca dos documentos de patente. O Apêndice A apresenta uma descrição mais detalhada da classificação IPC.

1.2 BASES DE DADOS DE PATENTES

Há décadas os pedidos de patentes e seus documentos correlacionados vêm sendo disponibilizados, possibilitando que sejam realizadas diversas pesquisas. Com a disponibilização dos dados de patentes das mais variadas formas, a limpeza e preparação destes dados tornam-se essenciais (WHALEN *et al.*, 2020).

Não só os textos dos pedidos de patentes estão sendo disponibilizados de forma mais acessível, mas seus metadados têm crescido bastante e vêm se tornando mais necessários para possibilitar a localização e organização dos documentos (WHALEN *et al.*, 2020).

Atualmente existem muitas bases de dados pagas e gratuitas com ferramentas de buscas associadas e que disponibilizam aos usuários os dados de pedidos de patentes. Algumas das bases de patentes mais importantes são: Espacenet[8], PatentView[9], Lens[10], CCD[11] e *GooglePatents*[12].

Normalmente, as bases de dados pagas apresentam algum diferencial para justificar o investimento das empresas que necessitam realizar buscas cada vez mais específicas. Entre estes diferenciais pode-se citar resumos feitos por especialistas e ferramentas de busca proprietárias. Em Abrantes *et al.* (2022) os autores realizaram um comparativo entre algumas das bases gratuitas, demonstrando diversos recursos que diferenciam uma base da outra.

---

[6] Site da versão em português do Brasil traduzida em grande parte por examinadores do INPI/BR. Disponível em: http://ipc.inpi.gov.br/ Acesso em: 4 out, 2024.

[7] Site da versão original na OMPI. Disponível em: https://ipcpub.wipo.int/ Acesso em: 4 out, 2024.

[8] Disponível em: https://worldwide.espacenet.com/ Acesso em: 4 out, 2024.

[9] Disponível em: https://patentsview.org/ Acesso em: 4 out, 2024.

[10] Disponível em: https://www.lens.org/lens/ Acesso em: 4 out, 2024.

[11] Disponível em: https://www.fiveipoffices.org/activities/globaldossier/ccd Acesso em: 4 out, 2024.

[12] Ferramenta do Google que disponibiliza informações sobre os pedidos de patentes de diversos países. Disponível em: https://patents.google.com/ Acesso em: 4 out, 2024.

1.3 CONCEITOS DE INTELIGÊNCIA ARTIFICIAL

Conforme pode-se ver na Figura 7, a inteligência artificial (*artificial intelligence*) é uma área do conhecimento que engloba aprendizado de máquina (*machine learning*) e aprendizado profundo (*deep learning*) (CHOLLET, 2021).

A inteligência artificial (IA) é uma área muito mais ampla do que o aprendizado de máquina e não necessita de uma etapa de aprendizado. A automação de tarefas utilizando regras codificadas já é considerado um algoritmo de IA e não necessita de aprendizado por parte da máquina (CHOLLET, 2021).

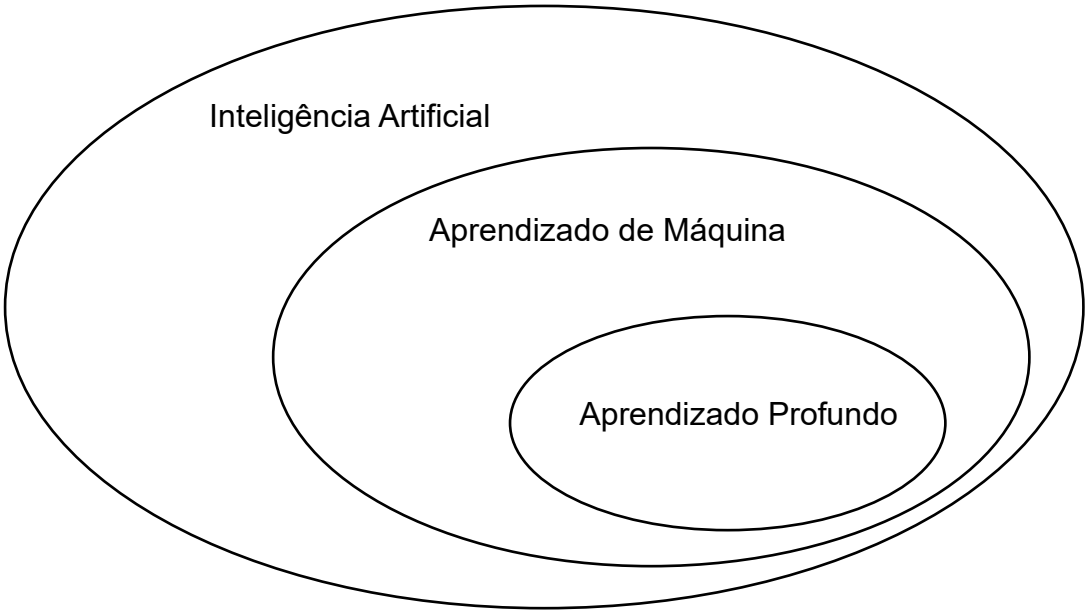


Figura 7 – Inteligência Artificial
Fonte: Chollet (2021)

A forma usual de solução de um problema por programação clássica consiste em implementar todas as regras possíveis, entrar com os dados e aguardar as respostas esperadas.

No aprendizado de máquina, essa ordem se inverte: o modelo recebe uma grande quantidade de dados juntamente com as respostas esperadas e as saídas são as regras estimadas pelas informações de entrada (CHOLLET, 2021).

A Figura 8 apresenta as entradas e as saídas das duas abordagens, a programação clássica e o aprendizado de máquina.

O aprendizado de máquina é um campo muito prático impulsionado por descobertas empíricas e muito dependente dos avanços de *software* e *hardware*. É basicamente uma disciplina de engenharia que utiliza diversos recursos da estatística trabalhando com volumes de dados grandes e complexos (CHOLLET, 2021).

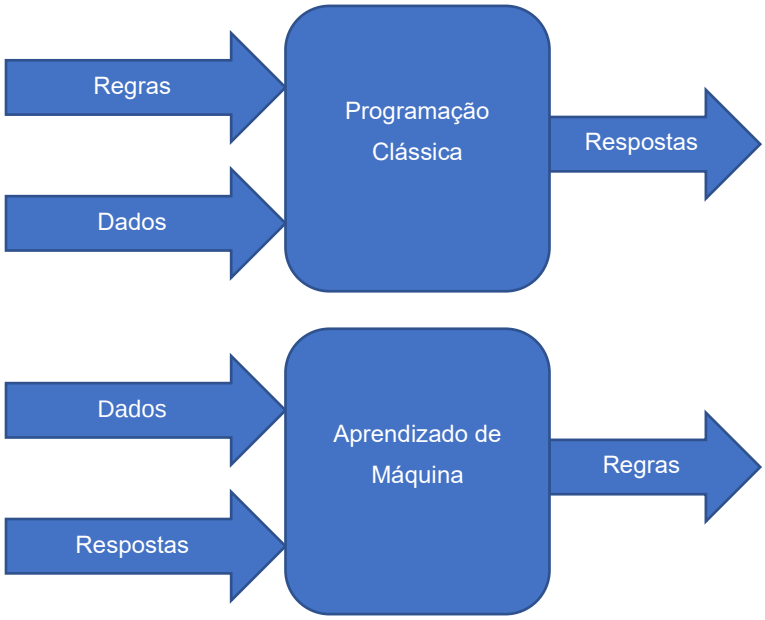


Figura 8 – Programação Clássica x Aprendizado de Máquina
Fonte: Chollet (2021)

O aprendizado profundo é um subcampo específico do aprendizado de máquina, sendo uma nova visão sobre representações de aprendizado a partir de dados que enfatizam o aprendizado de camadas sucessivas de representações cada vez mais significativas (CHOLLET, 2021). O termo profundo se deve a ideia de camadas sucessivas de representações das informações processadas.

A utilização de camadas de representação sucessivas originalmente foi associada as estruturas de conexões entre os neurônios do cérebro humano, inspirando diversos conceitos de aprendizado profundo. O termo rede neural refere-se à neurobiologia, mas na área de IA é associado aos modelos de aprendizado profundo. Neste trabalho, serão utilizadas ferramentas que utilizam os recursos de aprendizado profundo para a construção de modelos de linguagem.

1.4 SIMILARIDADE SEMÂNTICA

Na busca pela similaridade entre documentos de patentes, se faz necessário definir e entender melhor qual tipo de similaridade se deve explorar. Observando os textos de pedidos de patentes e procurando agrupar textos similares, verifica-se que estes normalmente são textos não estruturados. Mesmo as palavras utilizadas não seguem padrões rígidos, podendo um texto muito semelhante a outro em termos de assunto utilizar palavras diferentes. A análise sintática dos textos não foi considerada como uma boa opção, pois as estruturas dos textos podem ser distantes, porém próximos em termos de conhecimento.

Portanto, o caminho utilizado foi a análise semântica, ou seja, a análise do significado das entidades. Por exemplo, as palavras do idioma inglês *tea* e *coffee* (bebidas estimulantes) são mais semanticamente similares que as palavras *toffee* e *coffee*, as quais são sintaticamente semelhantes (HARISPE *et al.*, 2015, resumo).

Atualmente, as medidas de semântica são amplamente utilizadas em aplicações de Processamento de Linguagem Natural (PLN) ou *Natural Language Processing* (NLP) e tratamentos baseados em conhecimento, podendo tratar desde semelhanças entre palavra até semelhanças entre textos completos (HARISPE *et al.*, 2015, resumo).

Ao analisar textos não estruturados ou semiestruturados como textos simples, pode-se encontrar evidências da relação semântica entre as unidades de linguagem. Por exemplo, presume-se que as palavras café e açúcar são semanticamente mais próximas que as palavras café e gato. Portanto, é coerente concluir que a probabilidade da ocorrência de textos com as palavras café e açúcar é muito maior que a probabilidade do texto conter as palavras café e gato ao mesmo tempo (HARISPE *et al.*, 2015, p. 1).

As medidas semânticas são um assunto muito estudado pelos linguistas com o objetivo de comparar unidades de linguagem, tais como, palavras, frases, parágrafos e documentos. Alguns exemplos de aplicações que buscam comparar textos do ponto de vista do conceito são: verificação de plágio, geração de textos, geração de resumos, sistema de perguntas e respostas (HARISPE *et al.*, 2015, p. 2).

Muitas vezes, a estimativa de semelhança dos objetos comparados é criticada por diversos autores, pois dependendo das características comparadas tudo é semelhante. A noção de semelhança só faz sentido quando considerada dentro de uma representação parcial (HARISPE *et al.*, 2015, p. 5).

A seguir apresenta-se a definição de medidas semânticas, segundo Harispe *et al.,* (2015, p. 9):

> As medidas semânticas são ferramentas matemáticas utilizadas para estimar a força da relação semântica entre unidades de linguagem, conceitos ou instâncias, através de uma descrição (numérica) obtida de acordo com a comparação de informações que sustentam o seu significado.

Importante enfatizar que as medidas semânticas podem ser consideradas nos mais diversos domínios em um sentido matemático. Portanto, as comparações matemáticas podem ser entre palavras, frases, conceitos e/ou textos.

Considerando que todos os objetos podem ser comparados e que estes objetos possuem uma quantidade muito grande de características, observa-se cada vez mais a complexidade de comparar textos e buscar determinar a sua similaridade. Sendo assim, para aumentar a eficiência da determinação da similaridade entre documentos de pedidos de patentes foram necessárias condições de contorno, visando delimitar e organizar os documentos a serem comparados.

Neste trabalho, a classificação internacional de pedidos de patente (IPC) foi um ponto de partida relevante. Principalmente por ser um evento de intervenção humana no pedido que já tem a finalidade de agrupar pedidos que tratam de um mesmo assunto. Mesmo existindo a possibilidade de o julgamento realizado por um examinador apresentar falhas, a classificação atribuída ao pedido por uma pessoa com conhecimento técnico foi utilizada como ponto de partida.

1.5 MODELOS DE APRENDIZADO DE MÁQUINA

Um ponto fundamental em aprendizado de máquina é a questão entre otimização e generalização, buscando-se encontrar quais dados e suas respostas são suficientes para que se consiga treinar um modelo. Gerando regras genéricas e que consigam

interpretar entradas não utilizadas no treinamento, mas que geram saídas corretas. Isso seria um aprendizado de máquina ideal (CHOLLET, 2021).

Entretanto, por mais dados e respostas que se utilize no treinamento os modelos invariavelmente se tornam tendenciosos para os dados de treinamento, surgindo os problemas de *underfitting* e *overfitting* (ver item 1.5.2).

1.5.1 Épocas no treinamento

Muitas vezes, a quantidade de dados usados no processo de treinamento da rede neural é muito grande e precisa ser dividida em lotes. Cada lote inserido na rede neural constitui uma iteração. A cada iteração, a rede calculará os gradientes dos pesos em relação à perda do lote e atualizará os pesos dos parâmetros da rede neural. Quando a rede neural é submetida a todos os lotes do conjunto de dados de treinamento, tem-se uma época (CHOLLET, 2021).

Após o modelo ter sido submetido a todos os lotes do conjunto de treinamento, chega-se ao final de uma época. Então, o processo é reiniciado submetendo-se os dados ao modelo com os pesos ajustados pela época anterior, nos mesmos lotes de treinamento. Ao final dos lotes se terá uma nova época com novos pesos dos parâmetros ajustados. Esse procedimento é repetido por quantas épocas forem necessárias, para que se tenha um modelo com um desempenho considerado adequado.

A quantidade de épocas pode ser determinada com a utilização dos dados de teste e com a verificação do desempenho do modelo da rede neural. A cada época espera-se que o erro diminua. Entretanto, uma quantidade de épocas maior que o necessário irá gerar o problema de *overfitting* e, por outro lado, uma quantidade de épocas menor que o necessário ocorrerá o *underfitting*.

1.5.2 *Underfitting* e *overfitting*

Underfitting ocorre quando são executadas poucas épocas com os dados de treinamento e quando na etapa de testes o modelo gerado não acerta a resposta correta, nem mesmo para os dados de treinamento.

Overfitting ocorre quando um número exagerado de épocas é executado e o modelo basicamente tende a ficar certo somente para os dados de treinamento.

Em ambos os casos, a generalização do modelo não é alcançada.

1.5.3 Conjuntos de dados utilizados no modelo

Na busca pela otimização, generalização e tentativa de evitar *underfitting* e *overfitting* um recurso utilizado é a separação dos dados para os quais o modelo pode ser submetido em três conjuntos de dados: treinamento, validação e teste:

- Dados de treinamento – conjunto utilizado para geração dos coeficientes do modelo;
- Dados de validação – conjunto utilizado durante o treinamento para verificar se está ocorrendo o *overfitting* e determinar quando o treinamento deve parar;
- Dados de teste – conjunto não utilizado na etapa de treinamento. Visa verificar se o modelo é eficiente e se consegue um percentual razoável de acertos.

Diversos problemas ocorrem no modelo gerado, principalmente pela qualidade dos dados utilizados que apresentam as mais diversas formas de ruído. Por isso, quanto melhores os dados de treinamento, melhor será o modelo (CHOLLET, 2021).

O conjunto de dados de teste é utilizado somente no modelo já treinado, para que esse conjunto não influencie na geração dos coeficientes do modelo e o modelo possa ser avaliado de forma correta quanto ao seu desempenho.

A utilização de conjuntos de validação é útil para evitar problemas como *overfitting*. Este conjunto de validação é utilizado na etapa de treinamento apenas para monitorar a geração do modelo. Estes dados não influenciam diretamente nos cálculos dos coeficientes (GRANATYR, [*s. d.*]). Neste trabalho, não foi necessária a seleção de dados de validação, pois a técnica de tópicos utilizada já possui critérios de parada dos treinamentos dos modelos de linguagem.

1.6 PROCESSAMENTO DE LINGUAGEM NATURAL

O processamento de linguagem natural (PLN), mais conhecido pela expressão original em inglês *natural language processing* (NLP), é uma área ainda em desenvolvimento (PREMEBIDA, 2021).

Em Jacksi; Salih (2020), os autores realizam uma revisão sobre as ferramentas utilizadas para o agrupamento de documentos e concluem que a abordagem da similaridade semântica é a mais eficiente.

No presente trabalho, foi utilizada a abordagem semântica em uma técnica de modelagem de tópicos que permite a configuração dos modelos de linguagem, na qual podem ser utilizados diversos algoritmos de PLN existentes.

Considerando que os pedidos de patentes são baseados essencialmente em conteúdos técnicos, os tópicos gerados pelas técnicas de modelagem foram tratados como tópicos técnicos. Cada um destes tópicos técnicos foi representado por um conjunto de palavras. A associação de tópicos técnicos próximos aumenta a abrangência dos conteúdos técnicos abordados pelos documentos.

1.6.1 Inferência

Realizadas as etapas de treinamento dos modelos, os dados de teste podem ser submetidos para verificar se o modelo apresenta um bom desempenho; esta etapa chama-se de inferência. Neste procedimento qualquer documento pode ser submetido ao modelo treinado e o modelo irá retornar o tópico técnico, ao qual este documento melhor se encaixa.

Considerando o caso do INPI, que trabalha com pedidos de patente em sigilo, submeter estes pedidos ao modelo e verificar qual tópico o modelo de linguagem infere ao pedido é a operação ideal para tratar pedidos que estão nesta fase. Na inferência, os pedidos podem ser submetidos aos modelos sem que seus dados sejam expostos em ambientes acessíveis externamente. Assim, é possível encontrar documentos próximos dos pedidos em sigilo, sem que seus textos sejam utilizados diretamente em buscas que, muitas vezes, ocorrem em ambientes com pouca segurança. Por exemplo, a classificação dos pedidos de patentes é uma atividade

realizada durante o período de sigilo dos pedidos e o examinador de patentes precisa classificar o pedido sem expor os textos em sites não seguros.

1.6.2 Arquitetura de transformadores em PLN

Os transformadores ou *transformers* são arquiteturas em PLN. Os transformadores alteram uma sequência de entrada em uma sequência de saída. Basicamente, eles aprendem sobre o contexto e rastreiam as relações entre os componentes da sequência[13]. Uma característica desta arquitetura é que ela é unidirecional, ou seja, a arquitetura só verifica as palavras da esquerda para prever as próximas palavras da direita.

Por exemplo, considere esta sequência de entrada: "Qual é a cor do céu?" O modelo do transformador de linguagem usa uma representação matemática interna combinada ao conhecimento adquirido durante o seu pré-treinamento para identificar a relevância e a relação entre as palavras cor, céu e azul. Baseado neste conhecimento gera-se a resposta: "O céu é azul".

1.6.3 Arquitetura BERT em PLN

A arquitetura BERT – *Bidirectional Encoder Representations from Transformers*, (representações de codificador bidirecional de transformadores) é baseada na arquitetura *Transformers* (DEVLIN *et al.*, 2019).

Analisando cada uma das partes da arquitetura BERT:

- *Encoder Representations* – sistema de modelagem de linguagem, no qual insere-se um texto para a arquitetura e ela retorna a melhor representação numérica possível das palavras;
- *from Transformers* – indica que o *Transformer* define a arquitetura BERT;
- *Bidirectional* – indica que a arquitetura utiliza o contexto da direita e da esquerda das palavras no texto.

[13] Disponível em: https://aws.amazon.com/pt/what-is/transformers-in-artificial-intelligence/ Acesso em: 4 out, 2024.

1.6.4 Transformadores de sentenças

Os transformadores de sentenças (*Sentence Transformers* ou SBERT) são modelos pré-treinados, que podem ser usados para calcular incorporações ou pontuações de similaridade (THAKUR *et al.*, 2020)[14]. Alguns exemplos de aplicação são: pesquisa semântica, similaridade textual semântica e mineração de paráfrases. A paráfrase é um recurso linguístico que consiste em reformular um texto, mantendo o seu sentido original, mas utilizando outras palavras.

Atualmente, no repositório de transformadores de sentenças utilizado existem mais de 5000 modelos pré-treinados. Os modelos são treinados para finalidades específicas. Neste trabalho, foram utilizados quatro modelos originais all-MiniLM-L6-v2, all-MiniLM-L12-v2, all-distilroberta-v1 e all-mpnet-base-v2, os quais foram extensivamente avaliados em sua qualidade de incorporar sentenças e estes modelos são destinados para uso geral. Por outro lado, também foi escolhido o modelo LaBSE considerado eficiente para idiomas.

Os modelos all- * foram treinados em todos os dados de treinamento disponíveis (mais de 1 bilhão de pares de palavras de treinamento) e são projetados como modelos de propósito geral. Por exemplo, o modelo all-mpnet-base-v2 fornece melhor qualidade em relação aos outros modelos all- utilizados, enquanto all-MiniLM-L6-v2 é 5 vezes mais rápido do que o modelos all-mpnet-base-v2 e ainda oferece boa qualidade.

1.6.5 Corpus de texto

Em processamento de linguagem natural *corpus*[15] refere-se ao corpo do texto, que pode ser escrito ou falado em um ou mais idiomas. O plural de *corpus* é *corpora* e representa uma coleção de textos.

Um *corpus* pode ser definido como uma coleção de texto ou áudio originais organizados em conjuntos de dados. Ele pode ser constituído por textos de jornais, filmes, comentários e, para o presente trabalho, por textos que constituem os conteúdos técnicos dos pedidos de patentes.

[14] Disponível em: https://www.sbert.net/index.html Acesso em: 4 out, 2024.
[15] Corpo em latim.

A “Linguística de Corpus” é uma área da Linguística que se ocupa da coleta e análise de corpus. Utiliza ferramentas computacionais para realizar as análises dos textos coletados, buscando extrair informações relevantes para a realização de uma determinada tarefa (JOSÉ *et al.*, 2015).

Algumas das características que podem ser observadas em um *corpus* são[16]:

- Tamanho grande do *corpus*: Normalmente, quanto maior o tamanho de um corpus, melhor para treinar algoritmos projetados;
- Dados de alta qualidade: Devido à grande quantidade de conjuntos de dados, a qualidade dos dados precisa ser avaliada (AKKURT *et al.*, 2024), pois qualquer ruído nos dados pode gerar erros de grandes proporções;
- Dados limpos: A limpeza de dados (ver item 1.6.9) é um fator relevante para manter, ou mesmo melhorar, a qualidade dos dados;
- Equilíbrio: Um *corpus* de alta qualidade é um *corpus* equilibrado, no sentido de não apresentar dados que tornem os resultados tendenciosos. Por exemplo, se um pedido de patentes apresenta um número muito diferente de reivindicações do que os outros pedidos ele pode tornar os resultados tendenciosos. Por isso, a necessidade de decidir por algumas limitações em termos de quantidades.

1.6.6 TF-IDF

TF-IDF (*Term Frequency – Inverse Document Frequency*) é uma medida que busca quantizar a importância de um termo em um documento em uma coleção de documentos (corpus). TF-IDF é uma técnica muito utilizada em processamento de textos[17]. Os termos podem ser palavras, frases, etc. Os documentos podem ser frases, textos, etc. Neste trabalho, os termos serão as palavras das reivindicações e os documentos serão os textos das reivindicações de um pedido de patentes agrupados em um único conjunto.

A fórmula de TF-IDF é representada a seguir:

- $w_{i,j}$ – resultado de TF-IDF de i em j;
- i – termo (palavra das reivindicações);

[16] Disponível em: https://www.subex.com/article/corpus/ Acesso em: 4 out, 2024.
[17] Disponível em https://semantix.ai/tf-idf-entenda-o-que-e-frequency-inverse-document-frequency/ Acessado em: 4 out, 2024.

- j – documento (texto das reivindicações);
- N – número total de documentos;
- $tf_{i,j}$ – número de ocorrências de i em j pelo número de termos em j;
- df_i – número de documentos que contêm i.

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

TF refere-se à frequência de um termo (palavra das reivindicações) em um documento (texto das reivindicações). Quanto maior a frequência de uma palavra no texto, maior é sua relevância. A fórmula abaixo apresenta o cálculo de TF.

$$tf_{i,j} = \frac{número\ de\ vezes\ que\ um\ termo\ i\ ocorre\ em\ um\ documento\ j}{número\ total\ de\ termos\ em\ um\ documento\ j}$$

IDF refere-se à frequência inversa dos documentos. Isso indica que quanto maior o número de documentos que a palavra aparece, menor será a sua relevância. A fórmula abaixo apresenta o cálculo de IDF.

$$idf_i = \log\left(\frac{N}{df_i}\right)$$

A forma de cálculo de IDF pode variar, mas o ponto principal está em enfatizar as palavras que diferenciam um documento de outro. Por exemplo, uma palavra que está presente em todos os documentos do *corpus* não possui nenhuma relevância para diferenciar os documentos.

A técnica de modelagem de tópicos que foi utilizada neste trabalho utiliza uma variação do TF-IDF. Onde o TF-IDF foi ajustado para funcionar em um nível de tópico (*cluster*[18]) em vez de um nível de documento. Essa representação ajustada do TF-IDF

[18] *Cluster* é utilizado no sentido de agrupamento de documentos.

é chamada c-TF-IDF e leva em conta o que torna os documentos em um tópico diferentes dos documentos em outro tópico[19].

1.6.7 Redutores de dimensionalidade

A utilização de modelos de redes neurais busca processar as informações das amostras e transformar suas características em coeficientes. Estes coeficientes irão definir cada amostra. Normalmente, o número de coeficientes é muito grande e não pode ser apresentado de forma gráfica.

Os algoritmos redutores de dimensionalidade são utilizados com a finalidade de reduzir a quantidade de coeficientes que definem uma amostra. Transformando um conjunto de centenas de coeficientes em poucos coeficientes.

Quando as amostras são reduzidas a apenas dois ou três coeficientes, torna-se possível a visualização dos dados em um gráfico. Essa visualização possibilita aos pesquisadores um melhor entendimento do comportamento dos coeficientes calculados.

Neste trabalho, a técnica de redução de dimensionalidade utilizada foi o *Uniform Manifold Approximation and Projection*[20] (UMAP), que é uma técnica de aprendizado múltiplo para redução de dimensão. O UMAP é construído a partir de uma estrutura teórica baseada em geometria Riemanniana[21] e topologia algébrica (LIMA; MIRANDA, 2023). O resultado é um algoritmo prático escalável, que se aplica a dados do mundo real. O UMAP compete com o *t-distributed Stochastic Neighbor Embedding* (t-SNE) em termos de qualidade de visualização e preserva mais da estrutura global com desempenho de tempo de execução superior. O t-SNE é uma técnica não linear e não supervisionada[22] de redução de dimensionalidade, para exploração de dados e visualização de dados de alta dimensão[23].

---

[19] Disponível em https://maartengr.github.io/BERTopic/getting_started/ctfidf/ctfidf.html Acessado em: 5 out, 2024.

[20] Disponível em: https://github.com/lmcinnes/umap Acesso em: 4 out, 2024.

[21] Disponível em: https://mathworld.wolfram.com/RiemannianGeometry.html Acesso em: 4 out, 2024.

[22] Disponível em: https://www.ibm.com/br-pt/topics/unsupervised-learning Acesso em: 7 out, 2024.

[23] Disponível em: https://www.datacamp.com/pt/tutorial/introduction-t-sne Acesso em: 4 out, 2024.

O UMAP não tem restrições computacionais na incorporação de dimensão, tornando-o viável como uma técnica de redução de dimensão de propósito geral, para aprendizado de máquina (MCINNES; HEALY; MELVILLE, 2020).

1.6.8 Técnicas de agrupamento

Agrupamentos (*clustering*) é uma técnica de aprendizado de máquina, que divide dados em grupos, ou *clusters*, com base na similaridade. Ao juntar pontos de dados semelhantes e separar pontos diferentes em grupos, as técnicas buscam descobrir estruturas subjacentes em conjuntos de dados.

O algoritmo HDBSCAN[24] foi utilizado neste trabalho por ser o algoritmo indicado para agrupamentos na técnica de tópicos utilizada. HDBSCAN é um algoritmo de agrupamento que é projetado para descobrir grupos em conjuntos de dados com base na distribuição de densidade de pontos de dados. Diferente de outros métodos de agrupamento, o HDBSCAN não requer a definição do número de grupos e, portanto, ele é mais adaptável a diferentes conjuntos de dados. Ele usa regiões de alta densidade para identificar grupos e visualiza pontos isolados, ou de baixa densidade, como ruído. HDBSCAN é especialmente útil para conjuntos de dados com estruturas complexas ou densidades variadas, porque ele cria uma árvore hierárquica de grupos que permite ao usuário examinar os dados em diferentes níveis de granularidade[25].

1.6.9 Algoritmos de limpeza dos dados

Se por um lado, a capacidade de processamento dos equipamentos aumenta, por outro lado, a quantidade de dados disponível também aumenta. Com isso, os algoritmos de pré-processamento dos dados se mantém sendo necessários. A seguir, são apresentados os recursos de pré-processamento utilizados até o momento, mas conforme descrito na metodologia proposta, novos recursos devem ser avaliados sempre que possível.

[24] Disponível em: https://www.geeksforgeeks.org/hdbscan/ Acesso em: 4 out, 2024.
[25] Disponível em: https://www.geeksforgeeks.org/hdbscan/ Acesso em: 4 out, 2024.

1.6.9.1 Lematização (lemmatization) e estemização (stemming)

Lematização e estemização são duas técnicas de pré-processamento de texto que podem ser utilizadas individualmente ou em conjunto para tratar os textos e assim reduzir a carga computacional dos algoritmos de processamento de texto. Tanto a lematização quanto a estemização buscam reduzir uma palavra à sua raiz.

A lematização leva as palavras para suas formas fundamentais, retirando todas as inflexões, mas sempre mantendo uma palavra gramaticalmente correta. Na lematização, a classe gramatical da palavra é levada em consideração[26].

A estemização busca reduzir as palavras à sua raiz, mas diferente da lematização, ela só mantém a raiz da palavra, podendo o resultado perder o sentido.

A Tabela 2 apresenta exemplos das otimizações geradas em cada uma das técnicas.

Em Arts, Hou e Gomez, (2021), os autores utilizam entre outros recursos a lematização e estemização para verificar a criação e o impacto de novas tecnologias, analisando textos de pedidos de patentes.

Tabela 2 – Exemplos de lematização e estemização

Palavra original	Lematização	Estemização
amigos	amigo	amig
amigas	amigo	amig
amizade	amizade	amizad
carreira	carreira	carr
carreiras	carreira	carr

Fonte: Artigo no site Alura[27]

1.6.9.2 Técnica n-grama

N-grama são sequências contíguas de n itens de uma dada amostra de texto ou fala. Os itens podem ser fonemas, sílabas, letras ou palavras, de acordo com a aplicação. Eles são um método simples e eficaz para mineração de texto e tarefas de PLN, como:

[26] Disponível em: https://www.geeksforgeeks.org/python-lemmatization-with-nltk/ Acesso em: 5 out, 2024.

[27] Disponível em: https://www.alura.com.br/artigos/lemmatization-vs-stemming-quando-usar-cada-uma Acesso em: 5 out, 2024.

previsão de texto, correção ortográfica, modelagem de linguagem e classificação de texto[28].

Na análise de texto, se 'n' for 1, chama-se de unigrama; se 'n' for 2, será bigrama; se 'n' for 3, será trigrama, e assim por diante. Por exemplo, considerando a frase “n-grama é uma metodologia de mineração de textos”:

- Os bigramas seriam os seguintes pares “n-grama / é”, “é / uma”, “uma / metodologia”, “metodologia / de”, “de / mineração”, “mineração / de” e “de / textos”;
- Os trigramas seriam os seguintes trios “n-grama / é / uma”, “é / uma / metodologia”, “uma / metodologia / de”, “de / metodologia / de”, “metodologia / de / mineração”, “de / mineração / de” e “mineração / de / dados”.

O uso desse recurso no processamento dos textos faz com que as palavras sejam consideradas no contexto em que se apresentam e não mais isoladamente.

1.6.9.3 Palavras de parada (stop words)

As *stop words*, também conhecidas como palavras de parada, são termos que são frequentemente irrelevantes em um texto. Essas palavras são geralmente preposições, artigos, pronomes e outras palavras comuns que são usadas com frequência nos idiomas. A finalidade de identificar e retirar as *stop words* dos textos é reduzir o texto, a fim de manter as informações mais importantes.

Em Sarica e Luo (2021), os autores apresentam uma proposta para determinar *stop words* para as áreas técnicas, selecionando uma grande quantidade de pedidos de patentes e analisando os textos para encontrar as *stop words* relativas as áreas de conhecimento cobertos pelo sistema de classificação dos pedidos de patentes. Foram analisados aproximadamente 6 milhões de pedidos de patentes limitando-se em 500 palavras por documento e verificando-se as palavras irrelevantes em áreas pré-determinadas do conhecimento, ao final, obtendo-se uma lista de *stop words* específica para cada área do conhecimento escolhida.

[28] Disponível em: https://deepai.org/machine-learning-glossary-and-terms/n-gram Acesso em: 5 out, 2024.

1.6.10 Algoritmo de vetorização

Na modelagem de tópicos, a qualidade das representações de tópicos é essencial para sua interpretação. Não há uma maneira correta de criar representações de tópicos. Alguns casos de uso podem optar por n-gramas mais altos, enquanto outros podem se concentrar mais em palavras únicas sem nenhuma das *stop words*[29].

O algoritmo utilizado neste trabalho foi o CountVectorizer[30], que converte uma coleção de documentos de texto em uma matriz de palavras. Outra opção é o OnlineCountVectorizer, que é uma variante do CountVectorizer, com a inclusão de outros recursos.

1.6.11 Modelo de representação

Na técnica de modelagem de tópicos, a etapa de ajuste de representação busca selecionar palavras que possam melhor representar cada tópico (ACKERMAN, 2024).

O KeyBERT é uma técnica de extração de palavras-chave mínima, que aproveita incorporações de BERT para criar palavras-chave mais semelhantes a um documento[31].

1.7 TÉCNICA DE MODELAGEM DE TÓPICOS

BERTopic[32] é uma técnica de modelagem de tópicos que utiliza transformadores e realiza um ponderação com c-TF-IDF para criar tópicos (clusters) densos, permitindo tópicos facilmente interpretáveis, mantendo palavras importantes nas descrições dos tópicos (GROOTENDORST, 2022). A seguir serão descritas as características de cada etapa do BERTopic.

Em Naghshzan; Ratte 2023, os autores utilizam o BERTopic para gerar automaticamente resumos de manuais de instruções de funcionamento de aplicativos

[29] Disponível em: https://maartengr.github.io/BERTopic/getting_started/vectorizers/vectorizers.html Acesso em: 6 out, 2024.
[30] Disponível em: https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.CountVectorizer.html Acesso em: 6 out, 2024.
[31] Disponível em: https://github.com/MaartenGr/KeyBERT Acesso em: 6 out, 2024.
[32] Disponível em: https://maartengr.github.io/BERTopic/index.html Acesso em: 4 out, 2024.

a partir de informações disponíveis na internet. Eles aproveitaram as técnicas do BERTopic para extrair tópicos recorrentes e criar resumos concisos, demonstrando a flexibilidade da técnica de modelagem de tópicos por meio do BERTopic.

A Figura 9 demonstra a estrutura padrão de configuração do modelo BERTopic, detalhadas ao longo deste capítulo. Cada etapa trabalha de forma independente uma da outra. Por exemplo, a etapa de tokenização (tokenizer) não é diretamente influenciada pelo modelo de incorporação (embeddings), que foi usado para converter os documentos, o que nos permite flexibilidade na construção do modelo (GROOTENDORST, 2022).

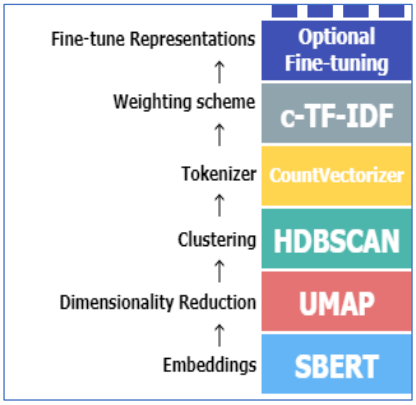


Figura 9 – Estrutura padrão do BERTopic
Fonte: (GROOTENDORST, 2022)

A Figura 10 apresenta diversas alternativas que podem ser utilizadas na construção do modelo BERTopic em cada uma de suas etapas.

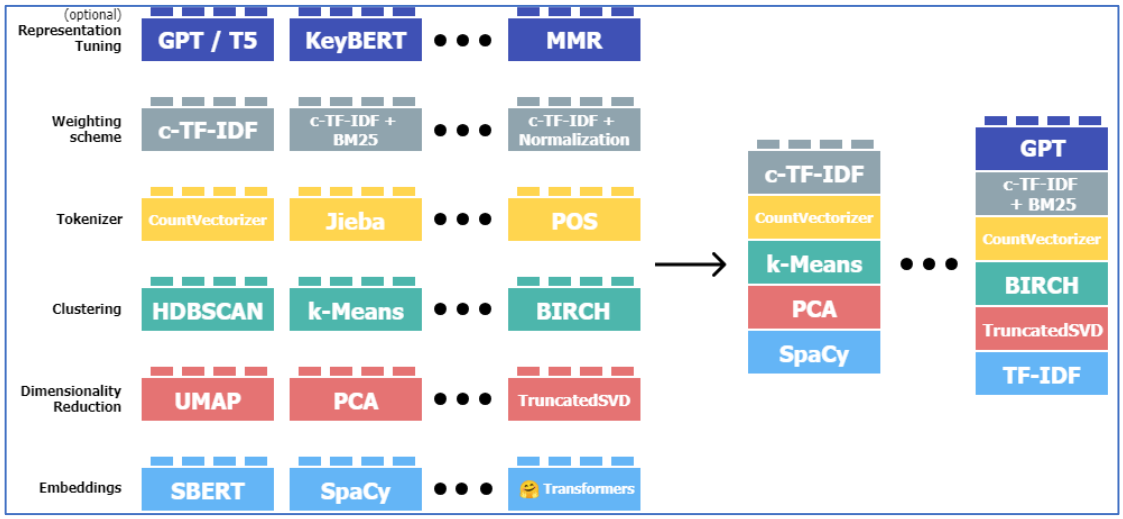


Figura 10 – Exemplos de alternativas para cada etapa
Fonte: (GROOTENDORST, 2022)

1.7.1 Incorporação (Embeddings)

A primeira etapa do BERTopic é a incorporação, na qual os documentos de entrada são transformados em representações numéricas. Embora existam outras técnicas para realizar a incorporação, normalmente são utilizados os transformadores de sentença[33] (por exemplo, all-MiniLM-L6-v2). Segundo Grootendorst (2022), os transformadores de sentenças são mais eficientes para capturar a semelhança semântica entre documentos do que outras técnicas de incorporação.

Não existe um modelo de incorporação perfeito. Portanto, a flexibilidade do BERTopic possibilita a realização de testes com diferentes modelos de incorporação, buscando comparar os desempenhos.

1.7.2 Redução de dimensionalidade (Dimensionality reduction)

A etapa de incorporação gera uma representação numérica de alta dimensionalidade. Em outras palavras, cada documento recebe um número muito grande de coeficientes numéricos para a sua representação. A etapa de redução de dimensionalidade busca reduzir a quantidade de coeficientes, mantendo apenas os coeficientes que melhor representam os documentos mapeados.

[33] Disponível em: https://www.sbert.net/docs/pretrained_models.html Acesso em: 4 out, 2024.

Uma solução é determinar de forma arbitrária a quantidade de coeficientes que se deseja para a representação dos documentos, observando aspectos como o custo computacional ou a possibilidade de representação gráfica. Este número de coeficientes depende de fatores como: capacidade computacional disponível, número de documentos utilizados em treinamentos e testes. Para que seja possível visualizar de forma gráfica a posição no espaço dos documentos e estimar as distâncias entre eles, o número de coeficientes precisa ser dois ou três.

O UMAP é usado como padrão no BERTopic, pois pode capturar o espaço local e global de alta dimensão, mesmo em dimensões inferiores (GROOTENDORST, 2022). Ou seja, mesmo com a redução de dimensionalidade, as distâncias originais entre documentos são preservadas de forma relevante.

1.7.3 Agrupamento (Clustering)

Após a redução de dimensionalidade, as incorporações semelhantes precisam ser agrupadas para que os tópicos possam se formar. Quanto mais preciso for o processo de agrupamento, melhor serão as representações de tópicos. No BERTopic, normalmente, o agrupamento é realizado por HDBSCAN (GROOTENDORST, 2022).

1.7.4 Vetorizadores (Vectorizer)

As diferentes unidades em que se pode separar os textos originais são chamadas de *tokens*. No caso das reivindicações, as separações foram em palavras e todas as palavras que restaram após os procedimentos de limpeza foram convertidas em *tokens*. Esse procedimento é chamado de tokenização. Os processos de vetorização de texto consistem em aplicar algum esquema de tokenização e, em seguida, associar índices numéricos aos *tokens* gerados (CHOLLET, 2021).

Esta etapa do BERTopic apresenta diversos recursos. Por exemplo: retirada de *stop words*, palavras mais frequentes e quantidades de tokens (palavras) para a representação dos tópicos.

1.7.5 Esquema de ponderação (Weighting scheme)

Essa etapa obtém uma representação precisa dos tópicos da matriz de palavras. O TF-IDF foi ajustado para funcionar em nível de tópico (cluster), em vez de nível de documento, esta representação TF-IDF ajustada é chamada c-TF-IDF (GROOTENDORST, 2022).

1.7.6 Ajuste de representação (Representation tuning)

Um dos principais componentes do BERTopic é sua representação e ponderação com c-TF-IDF. Este método pode gerar rapidamente uma série de palavras-chave para um tópico sem depender da tarefa de tópico (*cluster*). Como resultado, os tópicos podem ser atualizados de maneira fácil e rápida após o treinamento do modelo, sem a necessidade de treiná-lo novamente. Embora estes forneçam boas representações de tópicos, pode-se querer ajustar ainda mais as representações de tópicos (GROOTENDORST, 2022).

1.7.7 Medidas de coerência de tópicos

Coerência é uma medida do quão bem as palavras em um tópico relacionam-se entre si, refletindo a intuição humana do que torna um tópico coerente. Uma pontuação alta de coerência significa que o tópico é consistente, claro e relevante. Uma pontuação baixa de coerência significa que o tópico é vago, ruidoso ou irrelevante[34].

Em Mifrah 2020, Newman *et al.* 2010 e Rahimi *et al.* 2024, os autores realizam comparativos de coeficientes de coerência como métricas de escolha do melhor modelo de linguagem entre os disponíveis. Demonstrando a relevância do uso das medidas de coerência para a verificação da eficiência de cada modelo de linguagem em comparação com outros modelos de linguagem na realização de uma determinada atividade.

A coerência de tópicos é uma forma de julgar a qualidade dos tópicos por meio de um valor quantitativo e escalar. Neste trabalho, serão utilizadas duas das medidas de coerência *u_mass* (que normalmente apresenta valores entre -14 e 14) e *c_v* (que

[34] Disponível em: https://www.linkedin.com/advice/1/how-do-you-evaluate-coherence-perplexity Acesso em: 7 out, 2024.

apresenta valores entre 0 e 1). As duas medidas de coerência são valores de comparação entre os modelos treinados e o maior valor indica uma maior coerência[35].

A modelagem de tópicos é um método para localizar tópicos abstratos em uma grande coleção de documentos não rotulados e de maneira não supervisionada[36]. Com esse tipo de modelagem, é possível descobrir a mistura de temas ocultos, a qual varia de documento para documento em um determinado *corpus*. Portanto, a coerência das palavras em cada tópico é fundamental para garantir que os documentos no tópico sejam próximos.

1.8 FERRAMENTAS DE OUTROS ESCRITÓRIOS DE PATENTES

Muitos dos escritórios de patentes possuem ferramentas que realizam processamentos nos dados dos pedidos de patentes. Essas ferramentas são voltadas principalmente para a realização de buscas de documentos que auxiliem no exame dos pedidos de patentes.

Recentemente, o escritório do Reino Unido (UKIPO) lançou uma ferramenta de pesquisa que, segundo eles, ajudará os examinadores de patentes do UKIPO a realizar pesquisas de estado da técnica com maior profundidade, ajudando os examinadores a identificar citações potenciais durante o processo de exame de patentes de forma mais eficaz[37].

O escritório americano de patentes (USPTO) disponibiliza um serviço de pesquisas de patentes acessível ao público. Segundo o site, a ferramenta é um aplicativo de pesquisa de patentes baseado na *web*, que substituiu as ferramentas anteriores[38].

O escritório coreano disponibiliza a ferramenta *Korea Intellectual Property Information Search* (KIPRIS), a qual é um serviço abrangente e gratuito de pesquisa de informações sobre propriedade industrial disponível online na Coréia. Abrange informações de propriedade industrial, como patentes coreanas, modelos de utilidade,

---

[35] Disponível em: https://datascience.oneoffcoder.com/topic-modeling-gensim.html Acesso em: 7 out, 2024.

[36] Disponível em: https://www.ibm.com/br-pt/topics/unsupervised-learning Acesso em: 7 out, 2024.

[37] Disponível em: https://www.gov.uk/government/news/powerful-new-search-tool-will-help-ipo-maintain-patent-quality Acesso em: 4 out, 2024.

[38] Disponível em: https://www.uspto.gov/patents/search Acesso em: 4 out, 2024.

marcas registradas e design do Escritório Coreano de Propriedade Intelectual (KIPO) e outras informações de propriedade intelectual estrangeira. Além disso, o KIPRIS fornece serviço de recuperação de informações de administração de propriedade intelectual, que permite pesquisar exames, registros e status de testes[39].

Apesar da existência destas e outras ferramentas que podem ser utilizadas para agrupar pedidos semelhantes, os escritórios não disponibilizam informações mais detalhadas sobre os métodos utilizados. Então, os usuários destas ferramentas ficam limitados aos recursos existentes. Entende-se que estas ferramentas não são úteis para os objetivos do presente trabalho, mas podem ser relevantes para auxiliar na construção da base de dados em trabalhos futuros.

1.9 REFERÊNCIAS EM DESTAQUE

Neste item, são listadas algumas referências fundamentais para o melhor entendimento deste trabalho, indicadas como base para a realização de outros trabalhos semelhantes.

O livro Harispe *et al.* (2015), trata de similaridade semântica, apresentando os mais diversos aspectos e conceitos. A busca do significado de palavras e textos completos é tratada de forma clara e bem fundamentada.

O artigo Whalen *et al.* (2020), apresenta um exemplo muito completo de busca por similaridade entre documentos de patentes, incluindo desde os dados coletados até os resultados alcançados.

O artigo Hain *et al.* (2022), apresenta uma relação de abordagens utilizadas para encontrar patentes similares. Realiza uma crítica ao uso de textos para encontrar patentes semelhantes, argumentando que essa abordagem ainda não apresenta bons resultados. Argumenta, também, que o uso do resumo seria mais adequado para encontrar tecnologias semelhantes e que as reivindicações seriam mais adequadas para encontrar infrações entre patentes. Isso contribui com a decisão de utilizar as reivindicações para o presente trabalho, pois são as reivindicações que irão

[39] Disponível em: https://inspire.wipo.int/kipris Acesso em: 4 out, 2024.

determinar a proximidade entre as patentes e limitações dos objetos a serem explorados por seus detentores.

O artigo Younge; Kuhn (2016), apresenta uma abordagem de modelo de espaço vetorial, analisando o posicionamento dos documentos de patentes no espaço. Os autores alegam que a medida de similaridades baseada em textos funciona melhor que a classificação dos pedidos.

O artigo Mastrogiorgio; Gilsing (2016), se destaca por buscar tecnologias que a patente procura proteger utilizando outras tecnologias existentes. Esse trabalho demonstra que é importante mapear documentos similares e reforça as justificativas para melhorar as técnicas de detecção de similaridades entre documentos.

O artigo Grootendorst (2022), apresenta a construção do BERTopic, demonstrando diversas características que justificam o seu uso. Os autores construíram uma estrutura de suporte que permite o uso e o entendimento das ferramentas que estão em constante evolução. A modularização do BERTopic permite que diversas técnicas possam ser avaliadas de forma isolada em cada uma das suas seis etapas. A evolução nos algoritmos existentes e novos algoritmos pode ser utilizada e validada.

A presente tese se diferencia das referências, pois utiliza como balizador para a determinação de similaridade entre os documentos a correlação entre a patente e os documentos de família, citados e similares. Os documentos de patentes apresentam em sua grande maioria documentos de família, citados e similares os quais são correlacionados por especialistas nas mais diversas áreas do conhecimento.

2 METODOLOGIA

A metodologia proposta é a principal contribuição da tese, pois busca organizar procedimentos que possibilitem, dentro de um determinado contexto, agrupar documentos de patentes que apresentem similaridades em suas características semânticas.

A metodologia foi dividida em dois módulos, um primeiro módulo de treinamento e teste (MT) dedicado ao treinamento e teste dos modelos de linguagem e um segundo módulo de uso (MU) dedicado ao uso dos modelos de linguagem construídos e armazenados em um repositório de modelos de linguagem.

Os módulos foram estruturados em etapas e o presente capítulo descreve cada etapa de forma genérica, cabendo ao capítulo de desenvolvimento detalhar uma implementação, visando validar a metodologia.

Cada modelo de linguagem a ser construído utilizando essa metodologia se baseia em uma área do conhecimento específica. Conforme abordado anteriormente, a classificação IPC é uma das ferramentas utilizadas para organizar e identificar qual área do conhecimento é mais relevante em cada pedido de patente. Portanto, o ponto de partida de toda a metodologia foi a classificação internacional de patentes dos pedidos.

O módulo de treinamento e teste (MT) traz a figura do Pesquisador. O Pesquisador é uma pessoa ou pessoas que se preocupam em treinar, testar e disponibilizar modelos de linguagem voltados para uma determinada área do conhecimento. Este Pesquisador é o responsável pela análise e seleção dos dados dos pedidos que possam melhorar o desempenho dos modelos de linguagem. Do ponto de vista do processamento de linguagem natural trata-se do engenheiro de dados que pode trabalhar em conjunto com o cientista de dados.

No módulo de uso (MU) tem-se a figura do Usuário, o qual pode ser qualquer pessoa que possua documentos de uma área e que gostaria de verificar a proximidade entre estes documentos. Na estrutura do exame de pedidos de patentes o Usuário pode ser o chefe de divisão ou o examinador de patentes. O objetivo do Usuário é agrupar pedidos semelhantes dentro de um conjunto de pedidos de patentes que estão disponíveis.

Visando a integração entre o módulo de treinamento e teste (MT) e o módulo de uso (MU) existe a necessidade de um repositório por área do conhecimento de modelos de linguagem (RAML). Conforme surgem as necessidades do Usuário para uma determinada área do conhecimento o Pesquisador verifica a área, realiza o treinamento e disponibiliza no RAML o modelo de linguagem para usos futuros.

A Figura 11 apresenta a estrutura simplificada de toda a metodologia. No diagrama o Usuário verifica se existe um modelo de linguagem (ML) compatível com a necessidade. Se não existir solicita ao Pesquisador e se existir segue para o módulo de uso (MU). Por sua vez o Pesquisador recebe a solicitação do Usuário realiza os treinamentos e testes com o módulo de treinamento e teste (MT) e quando estiver pronto disponibiliza o ML no repositório RAML. A partir deste momento o modelo de linguagem fica disponível ao Usuário que fornece ao módulo de uso (MU) os números dos pedidos a serem agrupados e a quantidade de agrupamentos desejada.

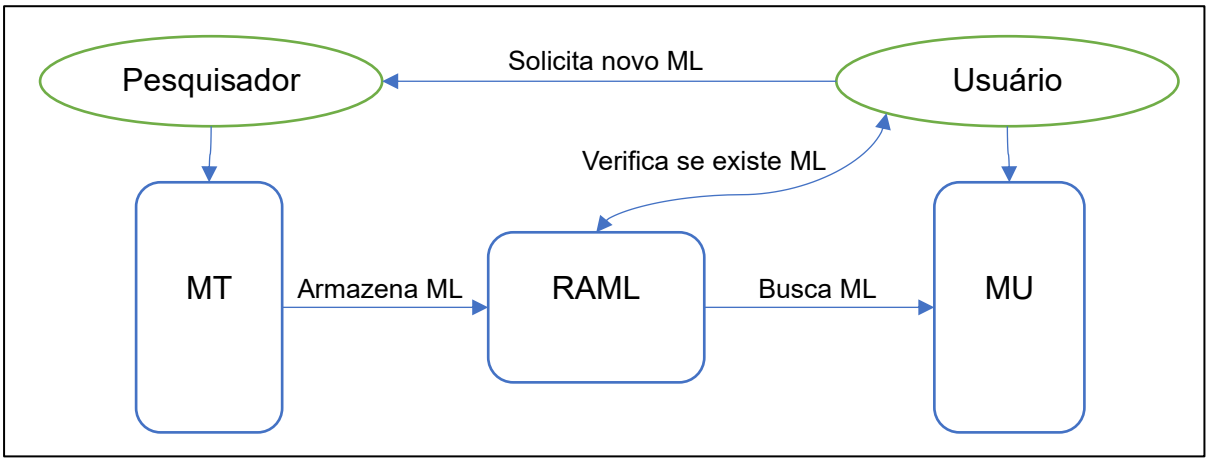


Figura 11 – Estrutura de funcionamento da metodologia
Fonte: Próprio autor

2.1 MÓDULO DE TREINAMENTO TESTE (MT)

O módulo de treinamento e teste (MT) dos modelos de linguagem foi dividido em nove etapas. As etapas de treinamento e teste estruturadas para a construção do modelo de linguagem foram nomeadas de T1 até T9, as quais serão descritas a seguir. Essa modularização visa facilitar o entendimento e possibilitar a implementação de melhorias em cada etapa de forma isolada.

2.1.1 Visão geral das etapas de MT

A Figura 12 apresenta um diagrama simplificado das etapas do módulo de treinamento e teste (MT) do modelo de linguagem. Logo a seguir, as etapas serão descritas de forma mais detalhada.

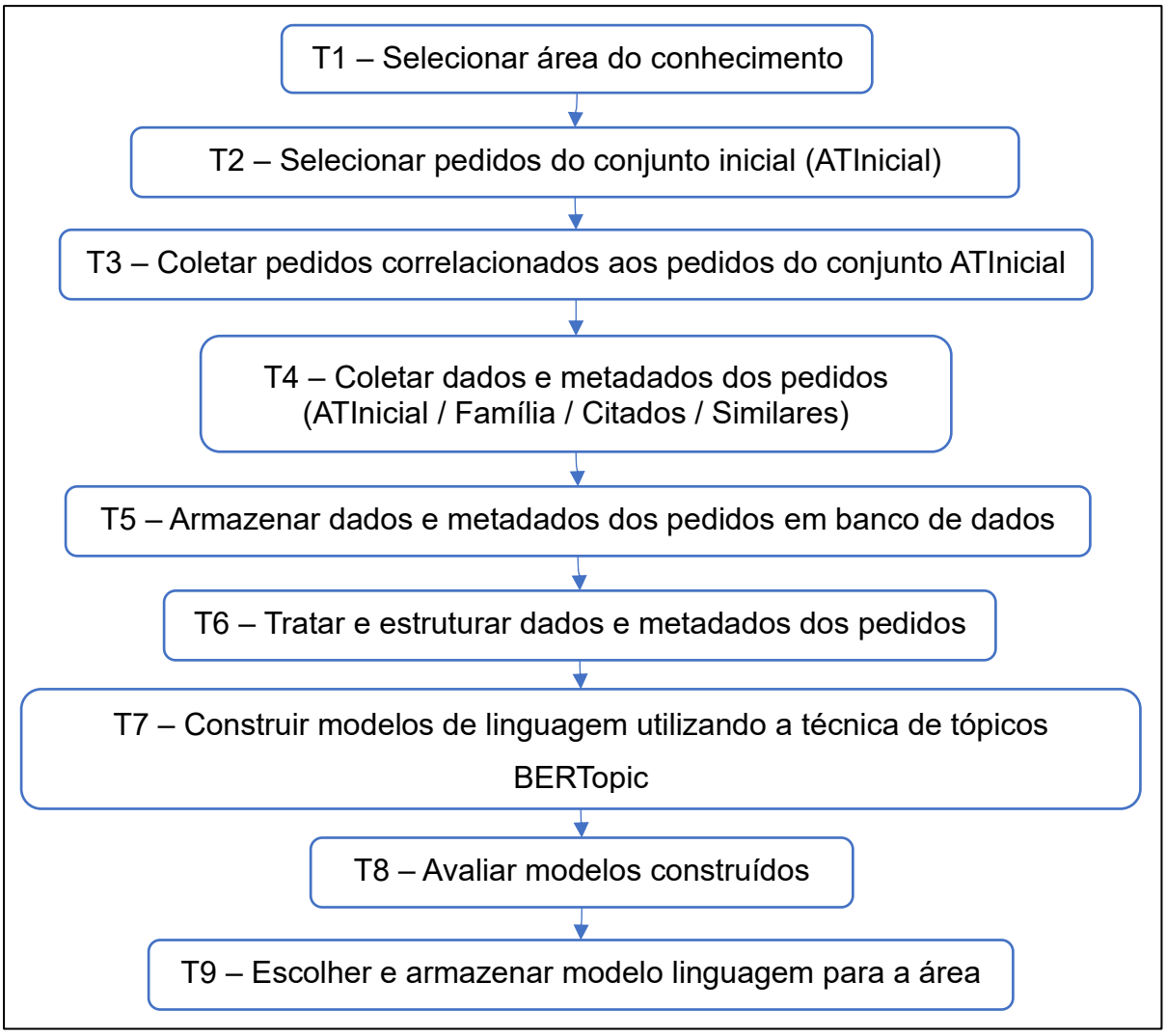


Figura 12 – Fluxo simplificado de treinamento do modelo de linguagem
Fonte: Próprio autor

2.1.2 Selecionar área do conhecimento (T1)

O Pesquisador decide para qual área de conhecimento o modelo de linguagem será direcionado, levando em consideração as solicitações do Usuário. Este modelo de linguagem será treinado e terá capacidade de realizar o agrupamento dos pedidos por tópicos. A abrangência do modelo é definida pela escolha do nível hierárquico do símbolo de classificação de patentes.

Se o Pesquisador definir o símbolo para o filtro dos pedidos como sendo a nível de classe na IPC, por exemplo G06, o conteúdo a ser analisado e separado em tópicos pelos algoritmos nos modelos será mais amplo do que escolher um símbolo em nível de grupo da IPC, por exemplo G06F16 (ver Apêndice A).

Por exemplo, no capítulo de desenvolvimento foi utilizado um símbolo de classificação em nível de grupo principal (G06F16/00) isso se deve a tentativa de limitar a área de conhecimento de forma mais restrita. Caso fosse utilizada uma seção inteira, por exemplo G, a quantidade de pedidos seria muito maior e os assuntos abordados seriam muito mais amplos. Em consequência, presume-se que os algoritmos teriam a tendência de agrupar pedidos por conceitos mais amplos.

Levando em consideração que os pedidos sejam agrupados por áreas do conhecimento específicas, resultados de agrupamentos por conceitos muito amplos reduziriam a eficiência do modelo de linguagem. Aqui, o Pesquisador precisa avaliar qual o nível hierárquico atende aos usuários finais do modelo de linguagem.

Importante ter em mente que o objetivo é buscar a similaridade semântica, ou seja, a similaridade de significado de um conteúdo com outro. Portanto, pedidos de áreas muito diferentes podem até apresentar uma grande quantidade de palavras iguais, mas o conjunto do pedido, muito provavelmente, terá um significado diferente. Isso justifica a utilização da classificação dos pedidos como um dos itens para limitação dos dados utilizados.

2.1.3 Selecionar pedidos do conjunto inicial (ATInicial) (T2)

O Pesquisador define parâmetros para delimitar os pedidos de treinamento e teste. Cada vez mais, a quantidade de documentos e informações vem aumentando e em termos de documentos de patentes a facilidade e possibilidades de buscar dados e metadados cresceu muito nos últimos anos. Existem muitas plataformas com bases de dados que disponibilizam informações sobre os pedidos de patentes e seus documentos relacionados, tanto de forma gratuita como paga, como por exemplo, GooglePatents, Espacenet, PatentScope, CCD, etc. A qualidade dos resultados está relacionada à utilização de bases de dados consolidadas e, por consequência, mais confiáveis.

O Pesquisador determina qual base de dados será a ideal para adquirir as informações dos pedidos e, dentro dos critérios de escolha, podem ser utilizados critérios como: quantidade de dados cadastrais disponíveis sobre os pedidos, informações personalizadas construídas pelos proprietários da base de dados, custos financeiros para utilização da base de dados, recursos de acesso às informações da base de dados, tipos de pedidos relacionados disponíveis na base de dados, etc.

Quanto mais bem estruturados são os dados da base e mais organizado for o armazenamento, mais fácil será para construir modelos para outras áreas do conhecimento, bem como mais eficiente e confiável será o processo de agrupamento dos pedidos em tópicos.

Escolhidos todos os filtros e critérios que o Pesquisador entender como necessários para a aquisição dos pedidos, ao final desta etapa, será obtida uma lista de pedidos. Tal lista é denominada, neste trabalho, de ATInicial, por se tratar de um conjunto de pedidos de uma amostra inicial, os quais serão a base para aquisição de pedidos correlacionados. São estes pedidos correlacionados que irão definir a eficiência dos modelos treinados.

A quantidade de pedidos no conjunto ATInicial só é limitada pela capacidade computacional em processar grandes quantidades de dados. Lembrando que para cada pedido do conjunto ATinicial serão inseridos os pedidos relacionados.

2.1.4 Coletar pedidos correlacionados aos pedidos do conjunto ATInicial (T3)

O Pesquisador define quais as informações serão coletadas dos pedidos do conjunto ATInicial e quais serão utilizadas, bem como quais possibilitam, de alguma forma, auxiliar na avaliação do desempenho do modelo como um todo.

A função fundamental do modelo resultante da metodologia é agrupar pedidos que sejam semelhantes uns aos outros. Diversas informações relacionadas aos pedidos podem ser utilizadas para verificar a assertividade do modelo.

Para possibilitar a geração de estatísticas de acertos dos modelos de linguagem verificou-se, durante o desenvolvimento da tese, a necessidade de encontrar uma relação entre os pedidos de patentes. Essa relação se torna mais legítima e confiável quando é gerada por especialistas em patentes ou por algoritmos robustos, o que

justifica o uso dos pedidos citados os quais são vinculados aos pedidos de patentes por especialistas da área ou por interessados e os pedidos similares os quais são vinculados aos pedidos por algoritmos das bases de dados.

No presente trabalho, optou-se por verificar a eficiência dos modelos, agrupando aos pedidos do conjunto inicial (ATInicial) os pedidos relacionados a estes em três categorias: pedidos da família, pedidos citados e pedidos similares.

Pedidos da família são pedidos que constam nas bases de dados por serem pedidos que pertencem às famílias dos pedidos das patentes do conjunto ATInicial. Estes pedidos apresentam uma relação próxima com o pedido que está sendo avaliado no momento.

Pedidos citados são pedidos que foram relacionados a um pedido do conjunto ATInicial em um determinado momento da vida do pedido. Podem ser pedidos citados por examinadores ou incluídos por outros interessados. Muitas vezes, são pedidos que foram utilizados como anterioridades ou que se encontravam no estado da técnica durante a análise do pedido. Normalmente, esse relacionamento é realizado por um especialista em patentes, o que torna essa categoria um parâmetro importante de verificação para o modelo de linguagem.

Pedidos similares são pedidos que algumas bases de dados, correlacionam, por meio de seus algoritmos, ao pedido em questão e o incluem na categoria de similares.

O Pesquisador precisa determinar um limite de pedidos para cada uma das categorias. Este limite deve ser definido de modo que o modelo não fique sujeito a distorções causadas por quantidades muito elevadas de pedidos relacionados a um mesmo pedido do conjunto ATInicial em uma das categorias.

2.1.5 Coletar dados e metadados dos pedidos (T4)

O Pesquisador define quais são as principais informações a serem coletadas e a forma de coletar os dados e os metadados dos pedidos do conjunto ATInicial (selecionado na etapa T3), bem como os dados e os metadados dos pedidos relacionados.

Todas as informações disponíveis podem ser coletadas e armazenadas, mas as informações que forem definidas como dados e metadados são essenciais para a realização das validações dos modelos de linguagem.

Cada base de dados de patentes apresenta uma estrutura diferente e formas de acesso que precisam ser analisadas. Deve ser definida qual técnica usar para coletar e organizar os dados dos pedidos e as interconexões entre os pedidos.

2.1.6 Armazenar dados e metadados dos pedidos em banco de dados (T5)

Os algoritmos de treinamento dos modelos de linguagem, tipicamente, são pesados computacionalmente. Portanto, as informações utilizadas por estes algoritmos precisam ser armazenadas da forma mais organizada possível.

O uso de bancos de dados locais e dedicados fazem com que a quantidade de dados seja apenas a necessária para a realização das atividades. Isso agiliza o processamento e facilita o entendimento da estrutura como um todo.

O Pesquisador precisa periodicamente reavaliar a forma de armazenamento de todos os dados e metadados em seu banco de dados. Eventualmente, modificações na estrutura de armazenamento podem ser relevantes.

2.1.7 Tratar e estruturar dados e metadados dos pedidos (T6)

O Pesquisador define quais procedimentos de limpeza das informações serão utilizados para tornar o sistema mais eficiente. A estruturação dos dados que serão submetidos ao modelo de linguagem também precisa ser definida. Normalmente, essa estrutura dos dados que é submetida aos modelos é denominada de *corpus*.

2.1.8 Construir modelos de linguagem utilizando a técnica de tópicos (T7)

O Pesquisador define quais parâmetros serão configurados no modelo de linguagem e a quantidade de modelos de linguagem que será construída para comparação.

Neste trabalho, a técnica de tópicos BERTopic (ver item 1.7) foi utilizada por apresentar uma gama muito grande de possíveis configurações. Aqui, serão adotados em sua maioria as configurações padrão da técnica, sendo realizadas variações no modelo de linguagem pré-treinado utilizado no módulo de incorporação do BERTopic.

2.1.9 Avaliar modelos de linguagem construídos (T8)

O Pesquisador busca encontrar o máximo de parâmetros para que se possa verificar o desempenho de cada modelo de linguagem construído na etapa T7 e comparar todos os modelos.

Diversos recursos para avaliação dos modelos de linguagem podem ser utilizados, por exemplo:

- hierarquia de tópicos – a estrutura hierárquica permite agrupar tópicos técnicos que apresentem proximidade, possibilitando reduzir o número de tópicos com poucos pedidos de patentes associados;
- mapas de calor – a observação dos mapas de calor possibilita ao Pesquisador localizar de forma visual regiões de alta ou baixa concentração de pedidos de patentes conforme o critério utilizado para a construção dos mapas;
- nuvens de palavras – permitem ao Pesquisador visualizar de forma rápida quais palavras se destacam em cada tópico, facilitando a retirada de palavras que não apresentam relevância para o tipo de assunto do tópico;
- erro médio quadrático – possibilita ao Pesquisador a visualização da variação de desempenho dos modelos de linguagem com a variação de um determinado parâmetro escolhido;
- coerência de tópicos – busca indicar a coerência entre as palavras que foram agrupadas nos tópicos conforme algum critério relativo de comparação, a coerência de tópicos pode ser muito útil, mas deve se ter sempre em mente que é uma métrica relativa e sempre deve ser feita a comparação em um mesmo contexto observando a variação de um parâmetro bem específico.

O desempenho dos modelos de linguagem com a utilização de pedidos de teste é a forma de avaliação comumente utilizada. Os modelos de linguagem são treinados com pedidos de treinamento e os pedidos de teste são submetidos ao modelo de linguagem já pronto. A resposta do modelo de linguagem aos pedidos de teste indica qual a eficiência do modelo de linguagem para pedidos desconhecidos no momento do treinamento.

2.1.10 Escolher modelo de linguagem para a área do conhecimento (T9)

Esta é a etapa de selecionar um dos modelos de linguagem, comparando o desempenho de todos os modelos de linguagem construídos. Após a escolha, podem ser realizadas outras análises para verificar o comportamento dos tópicos técnicos no modelo de linguagem selecionado. O modelo de linguagem escolhido é armazenado no repositório (RAML) para que o Usuário possa selecioná-lo quando for necessário por meio do módulo de uso (MU).

2.2 MÓDULO DE USO (MU)

O módulo de uso (MU) dos modelos de linguagem também foi organizado em nove etapas, nomeadas de U1 até U9. Com o objetivo de facilitar o entendimento, o módulo de uso (MU) foi construído com uma estrutura muito semelhante ao módulo de treinamento e teste (MT).

Em MT, o principal ator é o Pesquisador, o qual busca encontrar parâmetros para construir um modelo de linguagem que apresente o melhor desempenho possível para agrupar os pedidos de uma determinada área do conhecimento. Já em MU, o principal ator é o Usuário, o qual pode ser o chefe de divisão técnica ou mesmo o examinador de patentes. O objetivo do Usuário é agrupar pedidos de patentes que estão a sua disposição.

2.2.1 Visão geral das etapas de MU

A Figura 13 apresenta um diagrama simplificado das etapas do módulo de uso do modelo de linguagem. A seguir, as etapas serão descritas de forma mais detalhada.

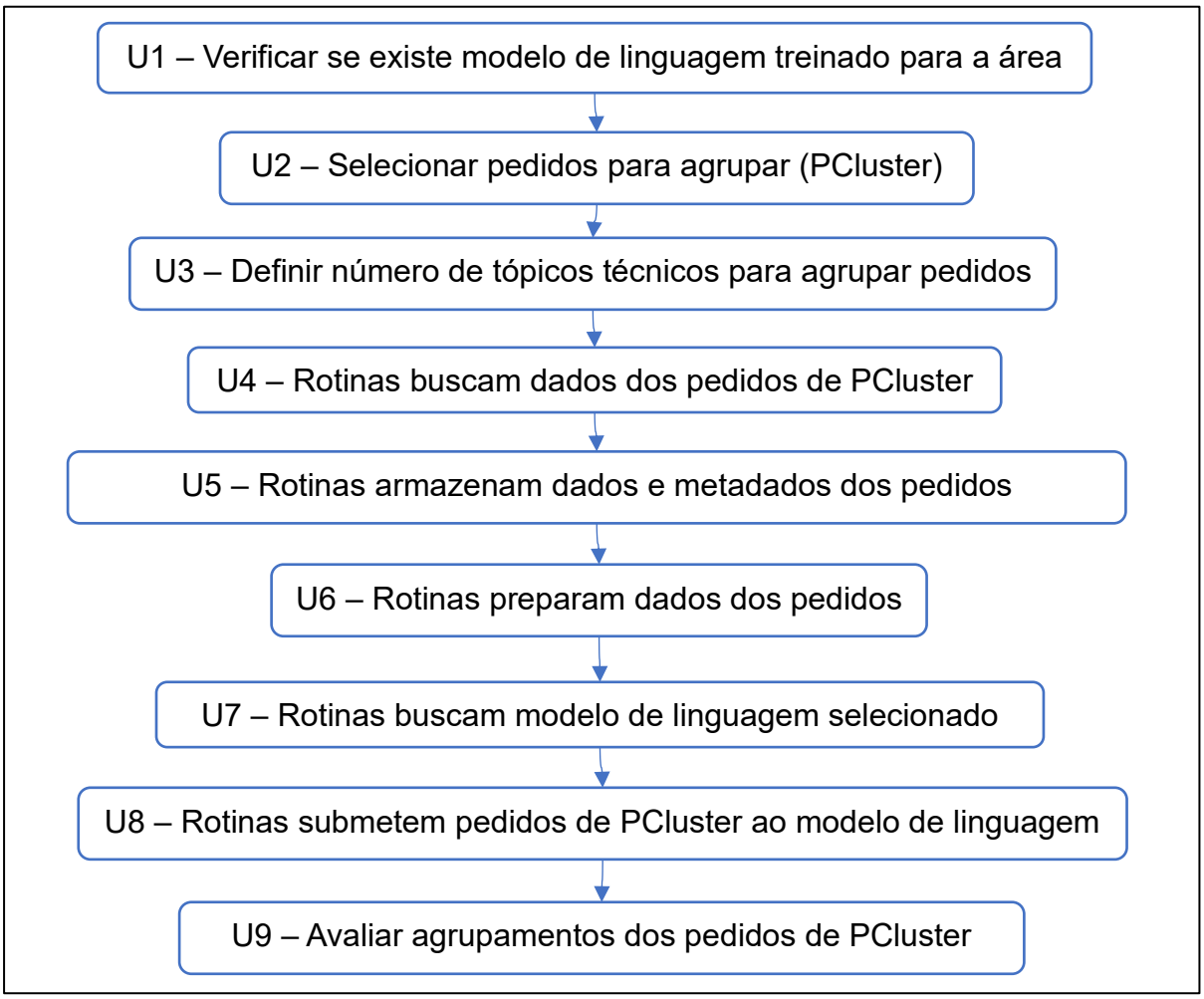


Figura 13 – Fluxo simplificado de uso do modelo de linguagem
Fonte: Próprio autor

2.2.2 Verificar se existe modelo de linguagem treinado para a área (U1)

O Usuário verifica se algum dos modelos de linguagem já treinados que estão no repositório (RAML) pode ser utilizado para os pedidos da sua área técnica. Caso não exista ainda nenhum modelo de linguagem que o Usuário considere adequado para a área dos pedidos, o Usuário precisa solicitar ao Pesquisador que seja realizado um processo de treinamento, especificando os símbolos de classificação.

2.2.3 Selecionar pedidos para agrupar (PCluster) (U2)

Nesta etapa, o Usuário informa os pedidos que precisam ser agrupados em tópicos técnicos para serem distribuídos aos examinadores. Esse conjunto foi denominado de

PCluster, por se tratar de um conjunto de pedidos que devem ser separados em grupos e pertencem à área escolhida em U1.

Pedidos de outras áreas que o Usuário considere próximas da área em que o modelo foi treinado também podem ser incluídos. Inclusive, a lista pode conter pedidos já examinados, para que o Usuário possa utilizar as anterioridades destes pedidos.

2.2.4 Definir número de tópicos técnicos para agrupar pedidos (U3)

Nesta etapa, o Usuário informa a quantidade máxima de tópicos técnicos que entende serem necessárias. Por exemplo, se o Usuário possui 200 pedidos em exame para agrupar e 20 examinadores, informar a quantidade de tópicos técnicos igual ao número de examinadores pode ser uma escolha inicial coerente. Certamente, com o resultado o Usuário precisa verificar se precisa aumentar ou diminuir essa quantidade de agrupamentos.

Os modelos de linguagem por tópicos costumam gerar uma quantidade elevada de tópicos. Tendo em vista que a quantidade de pedidos que o Usuário solicita para serem agrupados pode ser pequena, cada pedido poderia ficar em um tópico diferente, dificultando a identificação de agrupamentos.

Recebido o número máximo de tópicos técnicos desejados pelo Usuário, as rotinas calculam quais tópicos são mais próximos e agrupa os tópicos. Por consequência, agrupa os pedidos inferidos em relação a estes tópicos.

2.2.5 Rotinas buscam dados dos pedidos de PCluster (U4)

A partir dos números dos pedidos fornecidos pelo Usuário, as rotinas buscam os dados e metadados dos pedidos do conjunto PCluster. Estas informações coletadas dos pedidos são compatíveis com as informações definidas no momento de treinamento do modelo de linguagem. Por exemplo, se o modelo de linguagem foi treinado utilizando as reivindicações, os dados mínimos que precisam ser coletados são as reivindicações dos pedidos do conjunto PCluster. Caso os resumos tivessem sido utilizados seria mais coerente a coleta dos resumos.

2.2.6 Rotinas armazenam dados e metadados dos pedidos (U5)

As rotinas armazenam os dados e metadados dos pedidos do conjunto PCluster no banco de dados. Este armazenamento é importante para que não seja necessário acessar a base de dados externa se o Usuário incluir estes pedidos em novos conjuntos para agrupamento.

2.2.7 Rotinas preparam dados dos pedidos (U6)

As rotinas precisam realizar os mesmos procedimentos de limpeza das informações que foram executados para os dados de treinamento e teste. A estruturação dos dados que serão submetidos ao modelo de linguagem também deve ser semelhante à utilizada no treinamento e teste. Estes dados inferidos aos modelos também são denominados de *corpus*.

2.2.8 Rotinas buscam modelo de linguagem selecionado (U7)

Conforme a solicitação do Usuário, as rotinas buscam no repositório de modelos de linguagem RAML o modelo de linguagem escolhido pelo Usuário na etapa U1.

2.2.9 Rotinas submetem pedidos de PCluster ao modelo de linguagem (U8)

As rotinas submetem ao modelo de linguagem (já reduzido ao número de tópicos técnicos definido pelo Usuário na etapa U3) os dados dos pedidos do conjunto PCluster. O modelo de linguagem infere a cada pedido um dos tópicos técnicos disponíveis. Agrupa-se os pedidos de acordo com a quantidade de agrupamentos definida pelo Usuário em U3.

2.2.10 Avaliar agrupamentos dos pedidos de PCluster (U9)

Nesta etapa, o Usuário precisa verificar se o resultado do agrupamento dos pedidos foi satisfatório. Esta é uma etapa que só será possível ser realmente validada com o uso da estrutura toda em situações reais por um período. Não está no escopo deste trabalho, pois, para isso será necessária uma estrutura muito mais robusta do que a disponível para a validação da metodologia.

3 DESENVOLVIMENTO

A finalidade principal deste capítulo é validar a estrutura definida na metodologia, desenvolvendo um exemplo detalhado e analisando os resultados obtidos. Conforme definido na metodologia a estrutura proposta se divide em dois módulos: um módulo de treinamento e teste (MT) e um módulo de uso (MU) dos modelos de linguagem construídos.

Foram desenvolvidos três testes utilizando a metodologia proposta. O primeiro teste foi detalhado apresentando opções e especificações das etapas, tudo para possibilitar a validação da metodologia com a implementação da técnica de tópicos. Os outros dois testes descritos foram baseados neste teste detalhado, incluindo algumas modificações, para demonstrar a versatilidade da estrutura como um todo.

Conforme a estrutura proposta na Figura 11 apresentada na descrição inicial da metodologia todo o processo de treinamento se inicia de uma necessidade do Usuário em agrupar um conjunto de pedidos de patentes que pertencem a uma área do conhecimento. Para realizar a validação da metodologia foi necessário escolher uma área do conhecimento. Então, foi arbitrado o símbolo de classificação G06F16/00 para este primeiro teste, basicamente, por ser um símbolo de classificação que apresenta uma quantidade de pedidos suficiente para a realização dos testes.

3.1 LIMITAÇÕES ADOTADAS NA IMPLEMENTAÇÃO

A quantidade de dados bem como a quantidade de ferramentas que podem ser utilizadas para construir e testar os modelos de linguagem são muito diversificadas. Neste momento, o Pesquisador precisa decidir, em alguns casos de forma arbitrária, pela adoção de algumas condições de contorno. Muitas vezes é possível o Pesquisador realizar estatísticas com um conjunto dos dados e determinar os pontos de corte nos dados que serão adotados.

Neste capítulo de desenvolvimento foram adotadas as seguintes limitações:

- Foram utilizadas apenas as reivindicações como dados para treinamento e teste dos modelos de linguagem, os textos das reivindicações foram limitados as cinco primeiros reivindicações e um máximo de dois mil caracteres,

totalizando um máximo de dez mil caracteres relacionados a cada pedido utilizado nos treinamentos e testes;

- Classificações dos pedidos e datas (metadados) foram utilizados apenas para as filtragens dos dados;
- Foram escolhidos cinco transformadores de sentenças;
- Foi escolhido um período para realizar a busca dos pedidos do conjunto ATInicial, neste caso foram dois meses de pedidos US; e
- Foi determinado que os pedidos relacionados aos pedidos do conjunto ATInicial fossem limitados em 10 pedidos de família, 20 pedidos citados e 20 pedidos similares, totalizando um conjunto de no máximo 50 pedidos relacionados para cada pedido do conjunto ATInicial.

Neste primeiro momento estas limitações foram determinadas de forma arbitrária para uma construção inicial de todos os procedimentos. Entretanto, cada uma destas características poderia ser fundamentada com algum tipo de avaliação estatística.

3.2 ETAPAS DE MT – EXEMPLO DETALHADO

As etapas de T1 até T9 foram desenvolvidas obtendo-se, ao final, um modelo de linguagem treinado para a área do conhecimento escolhida na primeira etapa (T1).

3.2.1 Etapa T1 (selecionar área)

Foi escolhido para o teste o símbolo de classificação G06F16/00 cujo escopo é a recuperação de informação e estruturas de bases de dados ou arquivos para este fim. A opção de utilização de um símbolo de classificação em nível de grupo principal se deve à tentativa de limitar a área de conhecimento, de forma a gerar um modelo de linguagem eficiente para um assunto específico (ver item 2.1.2).

3.2.2 Etapa T2 (selecionar pedidos)

Conforme já apresentado no item 1.2 existem diversas bases de dados de pedidos de patentes sendo para essa etapa da metodologia o ponto mais relevante ser a

disponibilização de documentos correlacionados aos pedidos do conjunto ATInicial serem organizados nas categorias de família, citados e similares (ver item 2.1.3).

A presente tese utilizou a base de dados disponibilizada no site *GooglePatents*. A escolha desta base de dados foi realizada pelos seguintes motivos:

1. Disponibiliza um número muito elevado de pedidos de patentes dos mais diversos países. Por exemplo, documentos do Brasil (BR) estão disponíveis pouco mais de 1 milhão, dos Estados Unidos da América (US) mais de 21 milhões e documento da China (CN) são mais de 51 milhões;
2. Possui um serviço *web* que permite vários filtros, retornando os números dos pedidos que se encaixam nos parâmetros escolhidos;
3. Possibilita o *download* de arquivos com os números dos pedidos da seleção realizada;
4. Correlaciona de forma categorizada diversos documentos com um determinado pedido, apresentando seções de pedidos da família, pedidos citados e pedidos similares. Mesmo o *GooglePatents* não realizando referência ao tipo de família ser DOCDB ou INPADOC, percebe-se que os pedidos relacionados são baseados no conceito de INPADOC;
5. Possibilita a obtenção dos dados e dos metadados dos pedidos de patentes.

Na seleção do conjunto inicial de pedidos (ATInicial), além do símbolo de classificação G06F16/00, os documentos foram filtrados pelo país de depósito e pela data de depósito. A necessidade de limitação da quantidade de dados se deve basicamente às limitações de processamento.

Os pedidos do conjunto ATInicial foram definidos com as seguintes características:

- Símbolo: G06F 16/00;
- Pedidos de treinamento: data de depósito de 01/01/2020 até 29/02/2020 (dois meses de pedidos);
- Pedidos de teste: data de depósito de 01/03/2020 até 31/03/2020 (um mês de pedidos);
- País de depósito: US (Estados Unidos da América).

A Figura 14 apresenta a primeira tela da busca de um conjunto ATInicial de pedidos filtrados pelo símbolo G06F 16/00, o país com US e o período de depósito limitado no primeiro mês de 2020. Realizada esta busca, a página disponibiliza o *download* de um arquivo XLSX com informações sobre os pedidos encontrados. Uma característica observada na filtragem realizada no *GooglePatents* é que retornam todos os pedidos de patentes que possuem o símbolo filtrado incluindo os símbolos hierarquicamente inferiores a este símbolo.

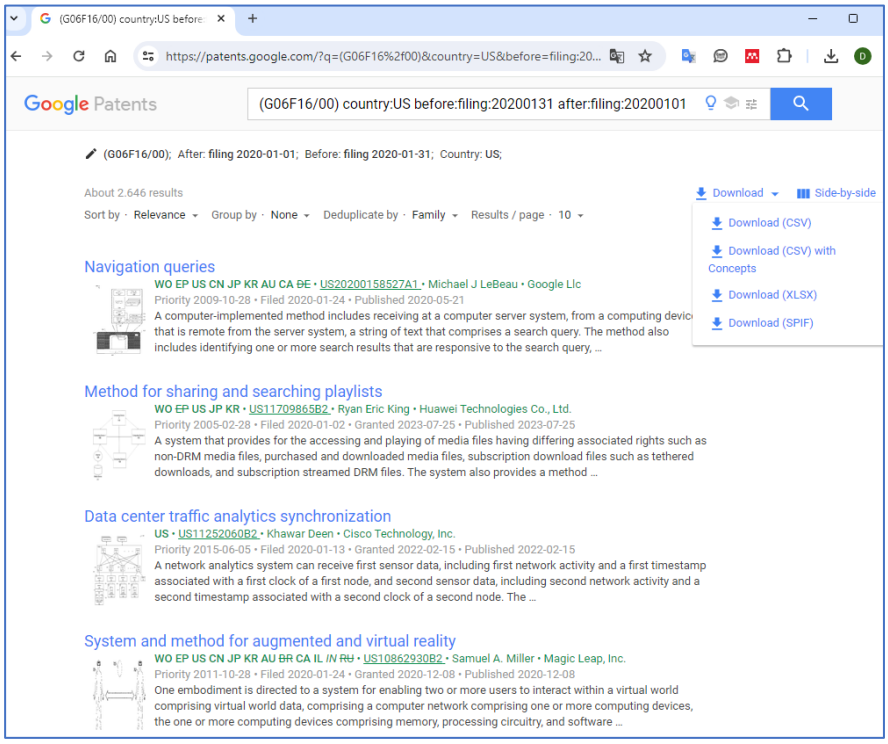


Figura 14 – Resultado da busca de um conjunto ATInicial
Fonte: *GooglePatents*

Ao final desta etapa foram obtidos dois arquivos XLSX: um com os pedidos para treinamento e outro com os pedidos para teste. Os pedidos relacionados nestes arquivos são os pedidos do conjunto ATInicial de treinamento e teste, respectivamente.

A Figura 15 resume as condições utilizadas para a aquisição dos pedidos dos conjuntos ATInicial de treinamento e teste. O resultado da pesquisa com estes filtros retornou para o conjunto ATInicial de treinamento uma quantidade de 2.701 pedidos de patentes e para o conjunto ATInicial de teste uma quantidade de 1.546 pedidos de patentes.

Figura 15 – Filtros para ATInicial
Fonte: Próprio autor

3.2.3 Etapa T3 (coletar pedidos correlacionados)

Esta etapa busca os pedidos de patentes do conjunto ATInicial constantes nos arquivos gerados ao final da etapa T2 e identifica os pedidos das categorias de família, citados e similares para todos os pedidos de ATInicial (ver item 2.1.4).

Conforme descreve a metodologia, é necessário limitar as quantidades dos documentos relacionados, pois alguns dos pedidos do conjunto ATInicial possuem centenas de documentos citados e isso poderia gerar distorções no treinamento dos modelos de linguagem. Também elevaria muito o número de documentos, o que geraria uma carga computacional muito grande e desequilibrada.

No sentido de equilibrar os dados dos pedidos relacionados aos pedidos do conjunto ATInicial, foi definido que os documentos de família seriam no máximo 10 e os citados e os similares seriam no máximo 20 cada. Sendo assim, cada pedido do conjunto ATInicial gerou, no máximo, 50 documentos que foram utilizados nos treinamentos e testes.

A Figura 16 apresenta os limites de pedidos determinados para cada categoria dos pedidos relacionados, em que k é o número de pedidos de patentes que retornaram no filtro dos pedidos do conjunto ATInicial. Como são dois conjuntos, tem-se o k_1 igual a 2.701 e k_2 igual a 1.546 que são as quantidades de pedidos resultantes da etapa T2. Ao final da etapa T3 foram coletados 55.281 pedidos de patentes correlacionados com os pedidos do conjunto ATInicial de treinamento e 31.184 pedidos de patentes correlacionados com os pedidos do conjunto ATInicial de teste.

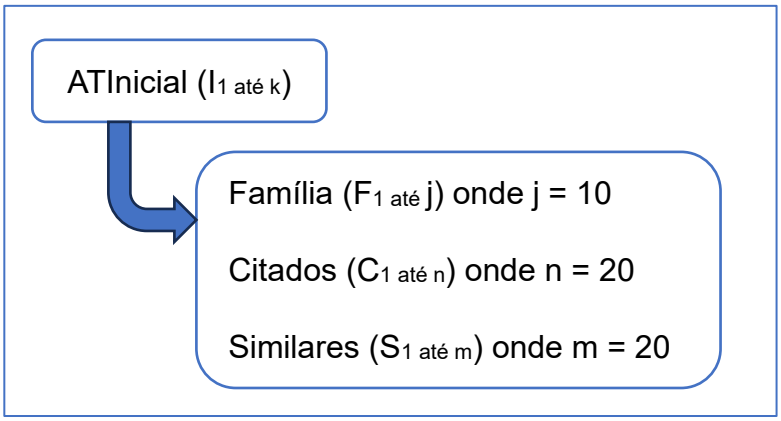


Figura 16 – Limites dos pedidos relacionados
Fonte: Próprio autor

3.2.4 Etapa T4 (coletar dados e metadados)

Foram definidos como dados os textos das reivindicações e como metadados os símbolos de classificação, data de depósito, país de depósito, números dos pedidos da família, números dos pedidos citados e números dos pedidos similares. A Figura 17 apresenta os conjuntos de pedidos pesquisados e as informações definidas como dados e metadados (ver item 2.1.5).

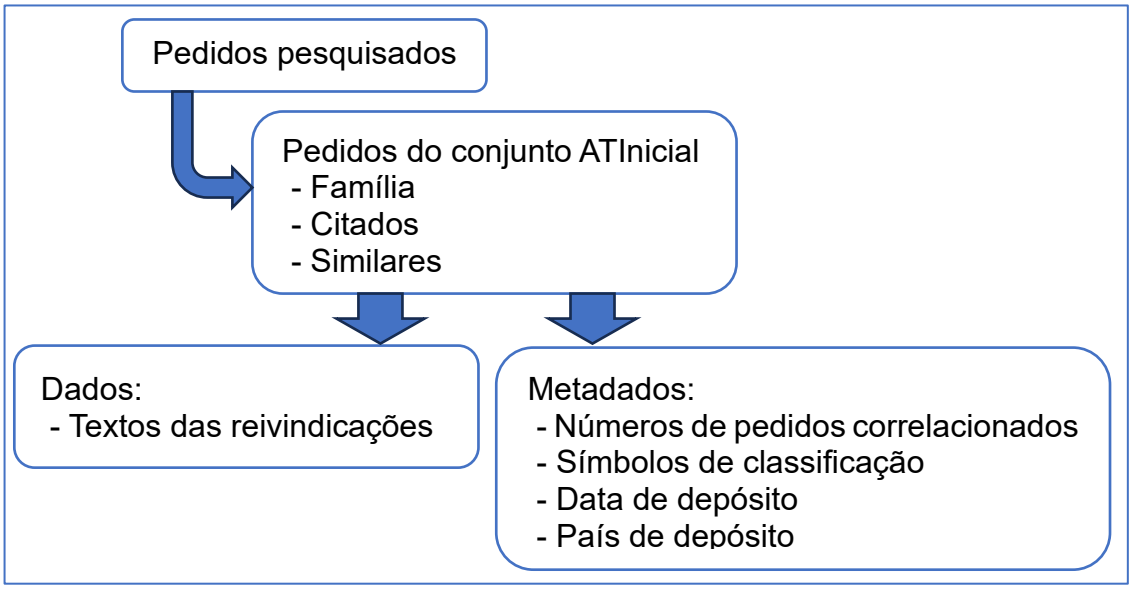


Figura 17 – Pedidos e informações coletados
Fonte: Próprio autor

A Figura 18 apresenta a página da base de dados *GooglePatents* do pedido US11709865B2, na qual as rotinas de coleta de dados encontraram as informações

do pedido de patentes pesquisado e os números dos pedidos de patentes correlacionados a este pedido.

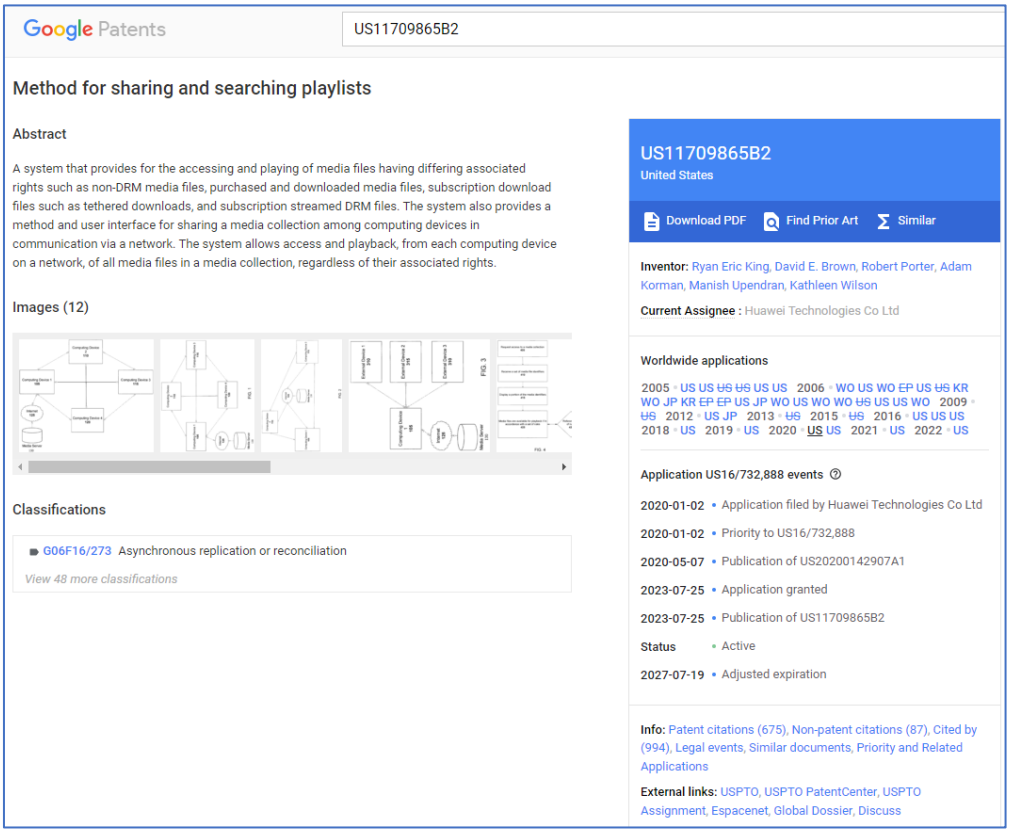


Figura 18 – Exemplo de página em que os dados são coletados
Fonte: *GooglePatents*

Conforme descrito anteriormente o Pesquisador precisa fazer limitações nos dados que serão utilizados para treinamento e teste, buscando equilibrar a base de dados para evitar distorções e respeitando a capacidade de processamento do seu ambiente de trabalho. Para os testes deste exemplo, os dados utilizados foram limitados os textos das cinco primeiras reivindicações dos pedidos e cada uma das reivindicações foi limitada em dois mil caracteres.

3.2.5 Etapa T5 (armazenar informações em banco de dados)

Utilizando as informações dos pedidos do conjunto ATInicial foram executadas rotinas de coleta de dados diretamente na base de dados do *GooglePatents*. Com os dados coletados, foram preenchidas as tabelas do banco de dados local. As principais informações coletadas no *GooglePatents* referente a cada pedido de patentes foram:

textos das reivindicações, símbolos de classificação, proprietários dos direitos, resumos, números de pedidos de família, citados e similares. Quanto mais organizados os dados são armazenados inicialmente, mais simples e eficientes se tornam as etapas seguintes de tratamento destes dados (ver item 2.1.6).

A Figura 19 apresenta a estrutura das tabelas, nas quais foram armazenados os dados de todos os pedidos pesquisados. Foram utilizadas cinco tabelas para o armazenamento dos dados adquiridos sobre os pedidos. Conforme descrito:

- Tabela gp_pedido contém informações básicas dos pedidos do conjunto ATInicial e dos seus pedidos de patentes correlacionados: família, citados e similares;
- Tabela gp_claim contém informações e textos das reivindicações dos pedidos;
- Tabela gp_cpc contém os símbolos de classificação dos pedidos;
- Tabela gp_assigned contém informações sobre os detentores dos direitos dos pedidos de patentes; e
- Tabela gp_public contém informações sobre os pedidos, conforme a sua categoria: família, citados e similares.

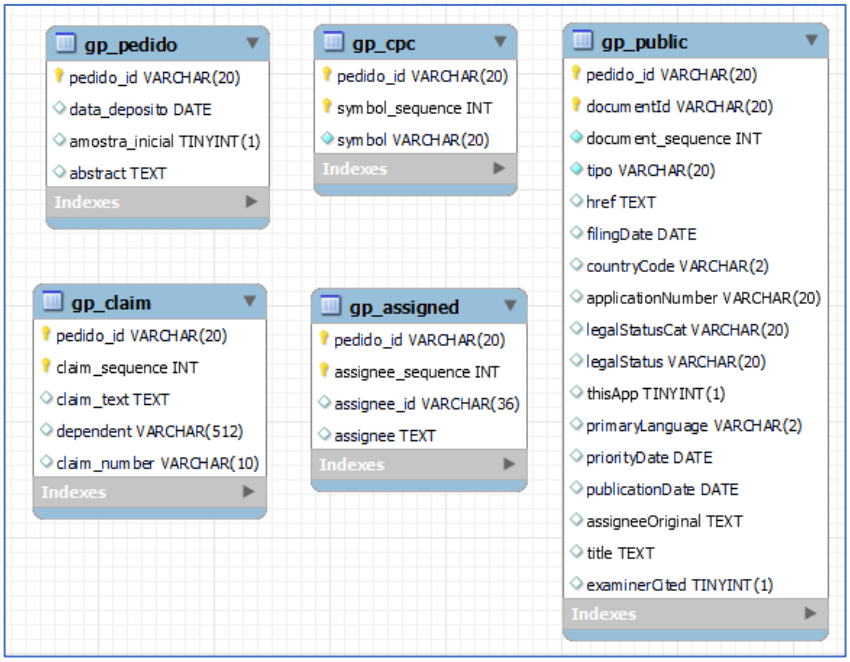


Figura 19 – Tabelas do banco de dados
Fonte: Próprio autor

A Figura 20 ilustra alguns registros, nos quais foram armazenados os resumos dos pedidos coletados. Analisando a base de dados o Pesquisador pode perceber alguma característica dos dados armazenados e realizar melhorias no desempenho do sistema retirando informações que podem gerar ruídos nos procedimentos seguintes. Por exemplo, a presença de caracteres especiais ou palavras em outros idiomas, as quais podem estar misturadas aos dados, podem ser percebidas pela observação dos dados.

pedido_id	data_deposito	amostra_inicial	abstract
US10642787B1	2020-01-23	1	In some embodiments, responsive to a user modifying a content of the file at a first client device (associated wi...
US10642993B1	2020-01-31	1	A method for sharing data in a multi-tenant database includes generating a share object in a first account comp...
US10654166B1	2020-02-18	1	Automation windows for attended or unattended robots are disclosed. A child session is created and hosted as ...
US10657684B1	2020-02-10	1	A system and method for displaying in a two-dimensional array the structured interaction of two variables movi...
US10664305B1	2020-01-17	1	Methods, systems, and apparatus, including computer programs encoded on computer storage media, for addi...
US10666778B1	2020-01-06	1	Methods, systems, and apparatus, including computer programs encoded on computer storage media, for perf...
US10672016B1	2020-02-11	1	There are disclosed marketing analytics apparatus and processes. There is a data store of interactions betwee...
US10678597B2	2020-01-15	1	Implementations of the present specification include receiving, from a client in a blockchain network, a request ...
US10678627B1	2020-01-24	1	Disclosed embodiments relate to automatically providing updates to at least one vehicle. Operations may includ...
US10691640B1	2020-02-13	1	Disclosed herein are computer-implemented methods; computer-implemented systems; and non-transitory, com...
US10693629B2	2020-01-09	1	Methods, systems, and apparatus, including computer programs encoded on computer storage media, for block...
US10693942B2	2020-01-06	1	Technologies related to resending hypertext transfer protocol (HTTP) requests are disclosed. One or more ope...
US10693958B2	2020-01-13	1	Methods, systems, and apparatus, including computer programs encoded on computer storage media, for addi...

Figura 20 – Exemplo de conteúdo da tabela gp_pedido
Fonte: Próprio autor

A Figura 21 apresenta alguns registros, nos quais foram armazenados os textos das reivindicações dos pedidos coletados, informando quais reivindicações são independentes e quais são dependentes.

pedido_id	claim_sequence	claim_text	dependent	claim_number
US10642787B1	1	1. A system comprising: a server system comprising one or more processors programmed with computer program instructions that, when ...	claim	1
US10642787B1	2	2. The system of claim 1, wherein, before the copy of the first file is transferred to the second client device, the transfer of the first met...	claim-dependent	2
US10642787B1	3	3. The system of claim 1, wherein the computer program instructions, when executed, cause the server system to: receive a copy of a s...	claim-dependent	3
US10642787B1	4	4. The system of claim 1, wherein the copy of the first file is automatically received from the first client device responsive to (i) a push re...	claim-dependent	4
US10642787B1	5	5. The system of claim 1, wherein the copy of the first file is automatically received from a first application at the first client device, and ...	claim-dependent	5
US10642787B1	6	6. The system of claim 5, wherein the first application is configured to create a first mobile object, and wherein the first mobile object is c...	claim-dependent	6
US10642787B1	7	7. The system of claim 6, wherein the first mobile object is configured to provide the copy of the first file to the proxy object, and wherei...	claim-dependent	7
US10642787B1	8	8. A method being implemented by a server system comprising one or more processors executing computer program instructions that, wh...	claim	8
US10642787B1	9	9. The method of claim 8, wherein, before the copy of the first file is transferred to the second client device, the transfer of the first met...	claim-dependent	9
US10642787B1	10	10. The method of claim 8, further comprising: receiving a copy of a second file from the second client device associated with the user, w...	claim-dependent	10
US10642787B1	11	11. The method of claim 8, wherein the copy of the first file is automatically received from the first client device responsive to (i) a push r...	claim-dependent	11
US10642787B1	12	12. The method of claim 8, wherein the copy of the first file is automatically received from a first application at the first client device, whe...	claim-dependent	12
US10642787B1	13	13. One or more non-transitory machine-readable media storing instructions that, when executed by one or more processors of a server ...	claim	13
US10642787B1	14	14. The media of claim 13, wherein, before the copy of the first file is transferred to the second client device, the transfer of the first me...	claim-dependent	14
US10642787B1	15	15. The media of claim 13, the operations further comprising: receiving a copy of a second file from the second client device associated w...	claim-dependent	15
US10642787B1	16	16. The media of claim 13, wherein the copy of the first file is automatically received from the first client device responsive to (i) a push r...	claim-dependent	16
US10642787B1	17	17. The media of claim 13, wherein the copy of the first file is automatically received from a first application at the first client device, whe...	claim-dependent	17
US10642993B1	1	1. A method comprising: granting to a second account of a multiple tenant database, access rights of a share object in a first account of ...	claim	1
US10642993B1	2	2. The method of claim 1, further comprising: granting the second account access rights to the alias object such that the second account ...	claim-dependent	2

Figura 21 – Exemplo de conteúdo da tabela gp_claim
Fonte: Próprio autor

A Figura 22 apresenta alguns registros dos símbolos de classificação de cada pedido coletado. Percebe-se que vários símbolos de classificação podem ser atribuídos para um mesmo pedido de patente.

pedido_id	symbol_sequence	symbol
US10642787B1	1	G06F16/122
US10642787B1	2	G06F15/16
US10642787B1	3	G06F16/00
US10642787B1	4	G06F16/10
US10642787B1	5	G06F16/128
US10642787B1	6	G06F16/13
US10642787B1	7	G06F16/14
US10642787B1	8	G06F16/176
US10642787B1	9	G06F16/178
US10642787B1	10	G06F16/182
US10642787B1	11	G06F16/1873
US10642787B1	12	H04L67/1095
US10642993B1	1	G06F21/6218
US10642993B1	2	G06F16/256
US10642993B1	3	G06F2221/2145
US10654166B1	1	G06Q10/103
US10654166B1	2	B25J9/1661

Figura 22 – Exemplo de conteúdo da tabela gp_cpc
Fonte: Próprio autor

A Figura 23 apresenta alguns registros dos detentores dos direitos dos pedidos de patentes coletados. Os direitos dos pedidos de patentes podem pertencer a uma ou mais empresas ou pessoas físicas.

pedido_id	assignee_sequence	assignee_id	assignee
US10642787B1	1		Topia Tech Inc
US10642993B1	1		Snowflake Inc
US10654166B1	1		UiPath Inc
US10657684B1	1		Effectivetalent Office LLC
US10664305B1	1		Advanced New Technologies Co Ltd
US10666778B1	1		Advanced New Technologies Co Ltd
US10672016B1	1		Impact Technologies Inc
US10678597B2	1		Lenovo PC International Ltd
US10678597B2	2		Advanced New Technologies Co Ltd
US10678627B1	1		Aurora Labs Ltd
US10691640B1	1		Advanced New Technologies Co Ltd
US10693629B2	1		Advanced New Technologies Co Ltd
US10693942B2	1		Advanced New Technologies Co Ltd
US10693958B2	1		Advanced New Technologies Co Ltd
US10698885B2	1		Advanced New Technologies Co Ltd
US10699076B2	1		Advanced New Technologies Co Ltd
US10700851B2	1		Advanced New Technologies Co Ltd
US10700852B2	1		Advanced New Technologies Co Ltd
US10706452B1	1		Capital One Services LLC
US10708060B2	1		Advanced New Technologies Co Ltd

Figura 23 – Exemplo de conteúdo da tabela gp_assignee
Fonte: Próprio autor

A Figura 24 apresenta alguns registros em que se pode perceber a categoria dos pedidos de patentes correlacionados.

pedido_id	documentId	document_sequence	tipo	href	filingDate	countryCode	applicationNumber	legalStatusCat	legalSt
US10642787B1	CN109313634B	150	Similares	/patent/CN109313634B/en	NULL	NULL	NULL	NULL	NULL
US10642787B1	CN109564529B	147	Similares	/patent/CN109564529B/en	NULL	NULL	NULL	NULL	NULL
US10642787B1	CN116010364A	158	Similares	/patent/CN116010364A/en	NULL	NULL	NULL	NULL	NULL
US10642787B1	CN116361778A	157	Similares	/patent/CN116361778A/en	NULL	NULL	NULL	NULL	NULL
US10642787B1	EP1130511A2	37	Citados	/patent/EP1130511A2/en	NULL	NULL	NULL	NULL	NULL
US10642787B1	JP2003528391A	152	Similares	/patent/JP2003528391A/en	NULL	NULL	NULL	NULL	NULL
US10642787B1	JP2006331439A	159	Similares	/patent/JP2006331439A/en	NULL	NULL	NULL	NULL	NULL
US10642787B1	KR101644666B1	148	Similares	/patent/KR101644666B1/en	NULL	NULL	NULL	NULL	NULL
US10642787B1	KR101647071B1	145	Similares	/patent/KR101647071B1/en	NULL	NULL	NULL	NULL	NULL
US10642787B1	US10067942B2	3	Familia	/patent/US10067942B2/en	2015-09-21	US	US14/860,289	active	Active
US10642787B1	US10289607B2	4	Familia	/patent/US10289607B2/en	2018-06-25	US	US16/017,348	active	Active
US10642787B1	US10530854B2	149	Similares	/patent/US10530854B2/en	NULL	NULL	NULL	NULL	NULL
US10642787B1	US10599671B2	142	Similares	/patent/US10599671B2/en	NULL	NULL	NULL	NULL	NULL
US10642787B1	US10642787B1	6	Familia	/patent/US10642787B1/en	2020-01-23	US	US16/750,399	active	Active
US10642787B1	US10754823B2	7	Familia	/patent/US10754823B2/en	2020-01-23	US	US16/750,435	active	Active
US10642787B1	US10805389B2	143	Similares	/patent/US10805389B2/en	NULL	NULL	NULL	NULL	NULL
US10642787B1	US10848557B2	144	Similares	/patent/US10848557B2/en	NULL	NULL	NULL	NULL	NULL
US10642787B1	US11003622B2	5	Familia	/patent/US11003622B2/en	2019-03-22	US	US16/361,641	active	Active
US10642787B1	US11061775B2	151	Similares	/patent/US11061775B2/en	NULL	NULL	NULL	NULL	NULL
US10642787B1	US11868800B1	156	Similares	/patent/US11868800B1/en	NULL	NULL	NULL	NULL	NULL

Figura 24 – Exemplo de conteúdo da tabela gp_public
Fonte: Próprio autor

Foram apresentados alguns registros das cinco tabelas utilizadas na base de dados de armazenamento local. Analisando estes dados o Pesquisador pode criar regras

para aumentar o desempenho dos modelos de linguagem construídos. Por exemplo, pedidos que foram classificados em vários símbolos de classificação e apenas um símbolo próximo ao símbolo que está sendo construído o modelo de linguagem, poderiam ser descartados para não gerar distorções no treinamento do modelo de linguagem.

3.2.6 Etapa T6 (tratar dados dos pedidos)

Todos os dados e metadados adquiridos na etapa T5 constituem o conjunto completo de informações armazenadas sobre os pedidos de patentes do conjunto ATInicial e seus pedidos correlacionados. Entretanto, é necessário selecionar e preparar o conjunto de dados que será submetido diretamente ao treinamento e teste dos modelos de linguagem construídos. Este conjunto é denominado *corpus* (ver item 2.1.7).

Optou-se por concatenar os textos das cinco primeiras reivindicações dos pedidos em um conjunto único relacionado a cada pedido. Então, foram concatenadas apenas as cinco primeiras reivindicações e limitou-se um máximo de 2000 caracteres por reivindicação. Isso foi necessário para que não se tivesse um conjunto muito elevado de dados relacionado com um mesmo pedido. Portanto, os textos das reivindicações dos pedidos foram de no máximo 10.000 caracteres (5 reivindicações x 2000 caracteres) para cada um dos pedidos utilizados.

Procedimentos de limpeza precisam ser realizados neste conjunto de textos para que os modelos de linguagem gerados sejam mais eficientes. Estes procedimentos irão ocasionar perdas de informações, entretanto, irão reduzir o custo computacional para o cálculo dos modelos de linguagem. As reduções dos dados disponíveis e os processamentos realizados nos textos restantes geram distorções, mas o Pesquisador tem a responsabilidade de buscar o melhor custo-benefício.

Neste primeiro teste, foram retiradas as *stop words* já disponibilizadas nas bibliotecas de tratamento de linguagem natural. Visto que todas as reivindicações utilizadas estão em inglês, a lista de *stop words* utilizada foi a da língua inglesa, uma vantagem por ser uma das listas mais consolidadas de todas. Pode-se incluir palavras novas em uma lista adicional de *stop words*, caso seja verificado que estas palavras são irrelevantes para o contexto de algum teste específico. Por exemplo, após a realização

de testes com pedidos de patentes, pode-se chegar à conclusão de que a palavra *method* está presente em quase todas as reivindicações dos pedidos. Isso, normalmente, torna essa palavra irrelevante na tomada de decisão dos modelos, sendo uma palavra que só aumenta a carga computacional do modelo.

Existem muitos outros procedimentos para a melhoria das informações utilizadas. Pode-se citar algumas técnicas que tendem a auxiliar na organização das informações, como: a lematização (busca reduzir as palavras a sua base, possibilitando o agrupamento de diversas palavras em uma única representação) e a estemização (busca a raiz das palavras reduzindo a quantidade de palavras) (ver item 1.6.9.1). Ao final do capítulo, nos dois testes adicionais estas técnicas foram abordadas apenas para demonstrar a flexibilidade da metodologia. Não foram realizados testes mais relevantes. O objetivo principal é validar a metodologia não sendo escopo do trabalho aprofundar pontos específicos.

3.2.7 Etapa T7 (construir modelo de linguagem)

A construção do modelo BERTopic é constituída de seis blocos. Estes blocos foram escolhidos conforme as mais diversas orientações de boas práticas fornecidas pelos desenvolvedores do BERTopic. Na abordagem realizada, apenas algumas modificações foram realizadas, mas devido à modularidade do modelo ainda existem muitos pontos que podem ser explorados (ver item 2.1.8).

Conforme visto anteriormente, os seis módulos do BERTopic são os seguintes:

- Incorporações – foram testados os seguintes transformadores de sentenças que segundo a literatura apresentam um bom desempenho no agrupamento de documentos similares semanticamente: all-MiniLM-L6-v2, all-MiniLM-L12-v2, LaBSE, all-distilroberta-v1 e all-mpnet-base-v2;
- Redutor de dimensionalidade – UMAP (definido por padrão no BERTopic);
- Agrupamento – HDBSCAN (definido por padrão no BERTopic);
- Tokenizador – CountVectorizer (definido por padrão no BERTopic) incluída a configuração de retirada de *stop words*;
- Esquema de ponderação – ClassTfidfTransformer (definido por padrão no BERTopic);

- Ajuste de representação – KeyBERTInspired com a configuração de número de palavras para definir a representação do tópico.

Cada uma das definições acima pode ser modificada para que se possa construir um modelo mais eficiente. Para realizar atividade com vários idiomas, existem incorporações multilíngue. Entretanto, optou-se por utilizar apenas textos em inglês.

3.2.8 Etapa T8 (avaliar modelos de linguagem)

Os pedidos de patentes para treinamento foram submetidos aos modelos construídos. Com base nestes modelos pré-treinados, foram gerados coeficientes influenciados por estes pedidos de treinamento. Este modelo retreinado é utilizado para inferir o tópico técnico ao qual novos pedidos de patentes melhor se encaixam (ver item 2.1.9).

Nesta etapa da metodologia o Pesquisador precisa gerar resultados baseados nos comportamentos dos modelos para que na próxima etapa seja possível selecionar o melhor modelo de linguagem construído.

Os pedidos selecionados para treinamento foram pedidos US na classificação G06F 16/00 com data de depósito de 01/01/2020 até 29/02/2020 e os pedidos selecionados para teste foram pedidos US com a mesma classificação e com data de depósito de 01/03/2020 até 31/03/2020. A Tabela 3 apresenta na coluna do conjunto ATInicial as quantidades de pedidos de patentes retornados pelas pesquisas realizadas no *GooglePatents*.

Tabela 3 – Quantidade de pedidos para treinamento e teste

Pedidos de patentes	ATInicial	Relacionados	Total
Treinamento	2.701	55.281	57.982
Teste	1.546	31.184	32.730

Fonte: Próprio autor

A coluna Relacionados da Tabela 3 lista os pedidos de patentes que apresentam relacionamentos de família, citados e similares com os pedidos da coluna do conjunto ATInicial. Os 2.701 pedidos de patentes do conjunto ATInicial para treinamento

originaram 55.281 pedidos relacionados, somando um total de 57.982 pedidos utilizados no treinamento.

A eficiência dos modelos de linguagem foi medida com base na assertividade do agrupamento, em um mesmo tópico técnico, dos pedidos do conjunto ATInicial com seus respectivos pedidos relacionados, os quais foram categorizados em família, citados e similares, listados na coluna Relacionados.

O procedimento para os 1.546 pedidos de patentes do conjunto ATInicial para teste foi o mesmo utilizado para os pedidos de treinamento. Os pedidos de teste foram submetidos ao modelo de linguagem somente após o treinamento, não influenciando nos cálculos dos coeficientes dos modelos de linguagem.

A Tabela 4 mostra que, destes 57.982 pedidos disponibilizados para treinamento, os modelos de linguagem utilizaram entre 37,42% e 60,83% dos pedidos para a realização do treinamento. Isso significa que parte dos pedidos submetidos aos treinamentos não foram agrupados em nenhum dos tópicos técnicos criados pelos modelos de linguagem. Estes pedidos que não foram relacionados a nenhum tópico técnico receberam o número “-1” (*outlier*) como tópico e foram descartados do treinamento. O descarte destes pedidos se deve ao algoritmo de treinamento verificar que eles poderiam gerar distorções aos modelos de linguagem.

Tabela 4 – Pedidos utilizados no treinamento

Transformadores	Pedidos para treinamento	Pedidos em tópicos técnicos	Aproveitamento
all-MiniLM-L6-v2	57.982	35.272	60,83%
all-MiniLM-L12-v2	57.982	33.615	57,97%
LaBSE	57.982	21.694	37,42%
all-distilroberta-v1	57.982	29.123	50,23%
all-mpnet-base-v2	57.982	33.374	57,56%

Fonte: Próprio autor

Os algoritmos de agrupamento, neste teste o HDBSCAN, analisam os pedidos submetidos ao treinamento e criam os tópicos. Quando os pedidos submetidos ao treinamento não se encaixam em nenhum dos tópicos criados estes pedidos são

classificados como *outlier*. Uma forma de reduzir os pedidos que foram classificados como *outlier* seria a melhoria na representação destes documentos modificando a representação c-TF-IDF.

3.2.8.1 Quantidade de tópicos técnicos gerados

Com o objetivo de verificar o comportamento dos modelos de linguagem construídos na etapa T8, os pedidos de patentes selecionados para treinamento e teste foram submetidos aos modelos em quantidades crescentes, até atingir o total de pedidos de patentes disponível para treinamento.

A Figura 25 apresenta a quantidade de tópicos técnicos gerada para cada modelo de linguagem, com a variação gradativa do número de pedidos de patentes utilizado para o treinamento. Todo o processo de treinamento foi realizado iniciando com 5 mil pedidos de patentes e aumentando 5 mil pedidos em cada teste.

Praticamente em todos os modelos de linguagem, percebe se um aumento gradativo na quantidade de tópicos técnicos. O gráfico da Figura 25 torna mais evidente que, com o aumento do número de pedidos para treinamento, mais tópicos técnicos são gerados. Portanto, mais assuntos diferentes são detectados pelos modelos de linguagem.

O modelo de linguagem com o transformador LaBSE apresentou uma certa instabilidade quando recebeu menos de 25.000 pedidos de patentes no treinamento. Contudo, quando o número de pedidos de patentes aumentou, ele já começou a ter a mesma tendência dos outros modelos de linguagem.

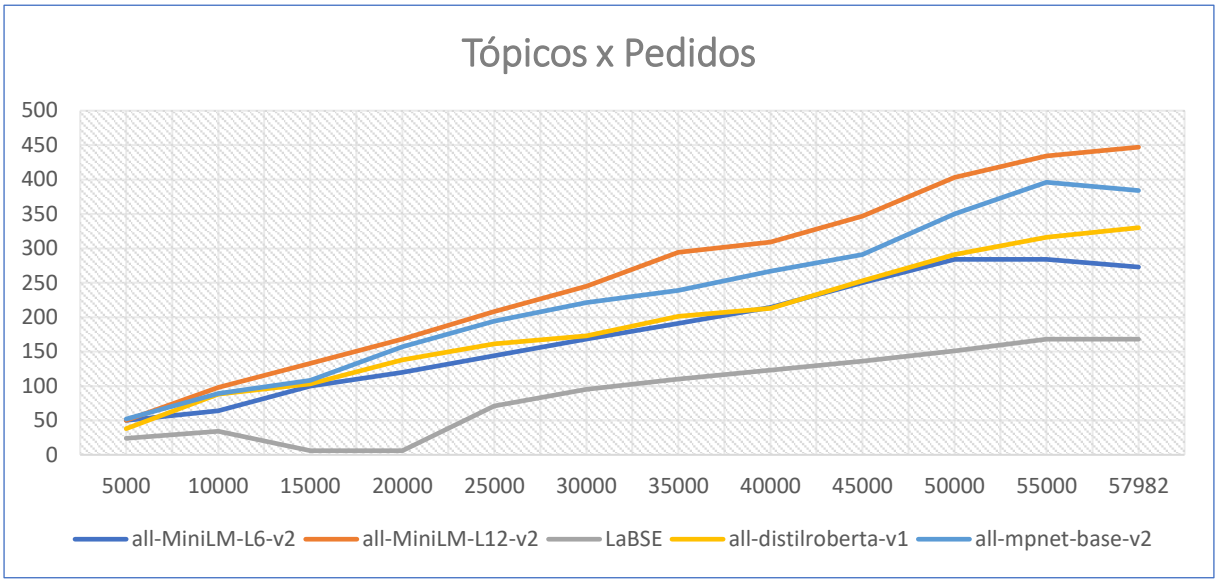


Figura 25 – Tópicos técnicos por pedidos
Fonte: Próprio autor

A quantidade de tópicos técnicos gerados pelos modelos de linguagem não é uma informação de grande relevância. A importância da Figura 25 está em demonstrar que o aumento de pedidos de patentes para treinamento gera um aumento na quantidade de tópicos. Esse comportamento indica que os dados utilizados apresentam alguma correlação.

3.2.8.2 Avaliação dos modelos de linguagem com RMSE

A raiz do erro quadrático médio (RMSE - *root mean-square error*) é uma métrica de avaliação utilizada em aprendizado de máquina para medir o desempenho de modelos de regressão (FILHO, 2023; OLUMIDE, 2023).

A fórmula a seguir apresenta a forma de cálculo utilizada para encontrar o RMSE dos modelos. Onde:

- n – número de tópicos técnicos do modelo de linguagem que apresentam pelo menos um dos pedidos do conjunto ATInicial;
- y_i – quantidade de pedidos de patentes correlacionados aos pedidos do conjunto ATInicial que estão no mesmo tópico i;
- p_i – quantidade de pedidos de patentes correlacionados aos pedidos do conjunto ATInicial que deveriam estar no mesmo tópico i;

- i – representação sequencial dos tópicos que contém pelo menos um dos pedidos do conjunto ATInicial.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - p_i)^2}$$

A Figura 26 apresenta o RMSE dos modelos de linguagem que apresentaram a tendência de unir os pedidos do conjunto ATInicial com os pedidos das suas respectivas famílias, confirmando o que seria esperado. Conforme o gráfico, o erro no agrupamento dos pedidos do conjunto ATInicial com os pedidos das famílias cai com o aumento do número de pedidos utilizados no treinamento. Com 57.982 pedidos (máximo de pedidos utilizados no treinamento), o erro cada vez mais se aproxima de zero. Isso representa entre 83,34% e 87,92% de acerto do modelo para estes agrupamentos nos tópicos técnicos. Este é um resultado relevante, pois significa que todos os modelos de linguagem aqui construídos têm tendência de convergir para agrupar os pedidos comprovadamente semelhantes. Possivelmente, com o aumento do número de pedidos submetidos ao treinamento, essa tendência de queda no erro deve continuar.

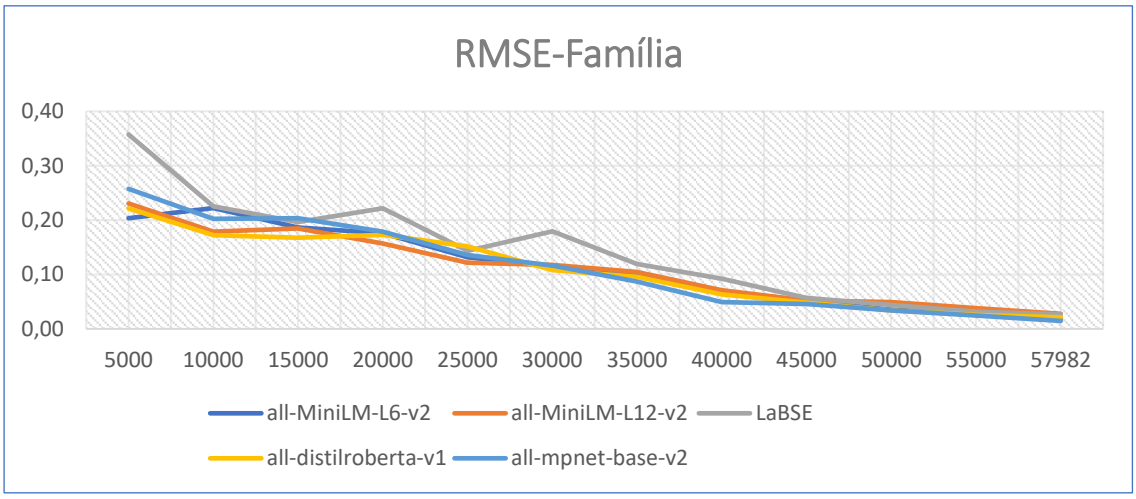


Figura 26 – RMSE-Família
Fonte: Próprio autor

A Figura 27 apresenta o RMSE da resposta dos modelos de linguagem para o agrupamento nos mesmos tópicos técnicos dos pedidos do conjunto ATInicial com os

pedidos citados. Os valores de RMSE no último ponto do gráfico para 57.982 pedidos (máximo de pedidos utilizado no treinamento) representam um percentual de acerto entre 14,54% e 21,84% conforme o modelo de linguagem. Normalmente, estes pedidos citados são relacionados aos pedidos em análise por especialistas na área do assunto do pedido e isso torna estes documentos relevantes.

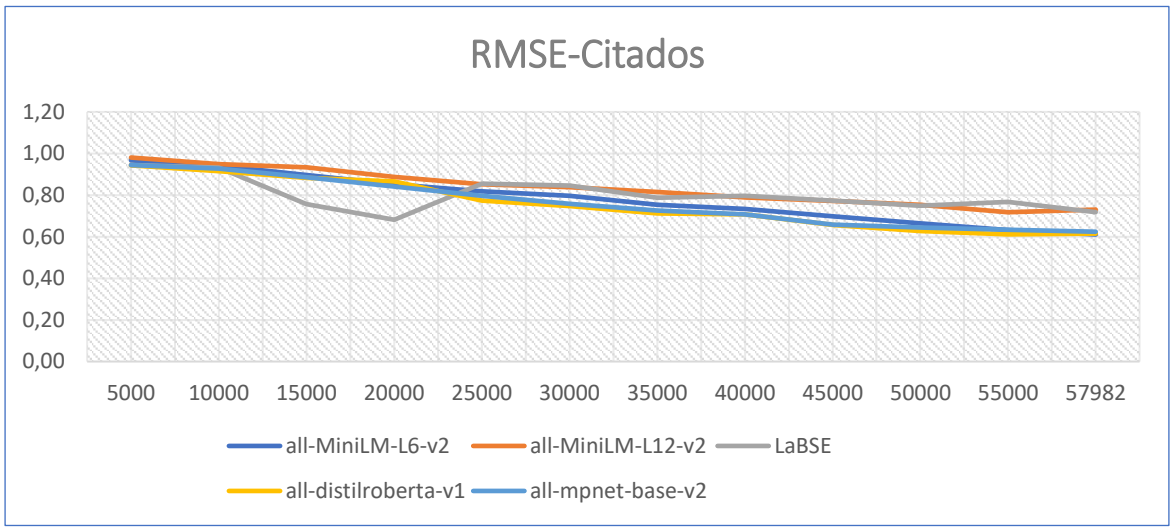


Figura 27 – RMSE-Citados
Fonte: Próprio autor

A Figura 28 apresenta RMSE da resposta dos modelos de linguagem para o agrupamento nos mesmos tópicos técnicos dos pedidos do conjunto ATInicial com os pedidos similares. Os valores de RMSE no último ponto do gráfico com 57.982 pedidos representam um percentual de acerto entre 18,32% e 25,68%, conforme o modelo de linguagem. São pedidos que foram considerados similares ao pedido pesquisado pelo site do *GooglePatents*.

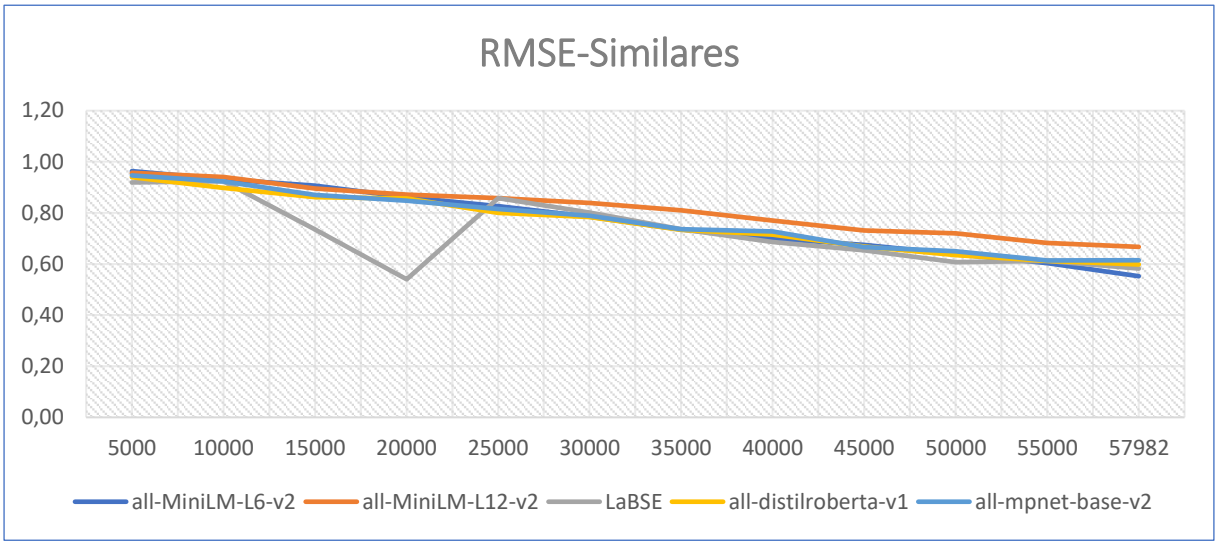


Figura 28 – RMSE-Similares
Fonte: Próprio autor

Não foram encontradas documentações que indiquem os procedimentos utilizados pelo site *GooglePatents* para justificar a relação dos pedidos em análise com os pedidos indicados como similares. Entretanto, observando-se a semelhança entre os percentuais de acerto dos pedidos citados (entre 14,54% e 21,84%) e dos pedidos similares (entre 18,32% e 25,68%), isso pode indicar que as duas categorias apresentam um comportamento equivalente.

O RMSE dos pedidos citados e similares ser maior que o RMSE dos pedidos da família indica que os modelos de linguagem são sensíveis às diferenças entre os documentos, aproximando documentos do conjunto ATInicial com documentos das suas famílias e distanciando dos documentos citados e similares.

A tendência de o RMSE ser menor para os pedidos similares do que para os pedidos citados pode ser atribuída a forma como essa correlação foi realizada. Os pedidos citados foram correlacionados por especialistas e interessados e os pedidos similares foram correlacionados por algoritmos de forma semelhante aos procedimentos realizados aqui.

3.2.8.3 Verificação da coerência dos tópicos nos modelos

Conforme visto no item 1.7.7 as medidas de coerência de tópicos são medidas relativas que permitem a realização de comparativos entre os modelos de linguagem

de tópicos que estão sendo avaliados. Quanto maior o número encontrado, mais coerentes são as palavras nos tópicos. O coeficiente *c_v* pode ter valores entre 0 e 1 e o coeficiente *u_mass* pode ir de -14 até 14.

A Tabela 5 apresenta as coerências *c_v* e *u_mass* dos cinco modelos de linguagem treinados. Verifica-se que o modelo LaBSE apresenta uma coerência *c_v* menor que os demais. Os outros quatro modelos apresentam basicamente a mesma coerência *c_v*. Verifica-se que todos os modelos apresentaram uma coerência *u_mass* muito semelhante.

Tabela 5 – Medidas de coerência

Transformadores	c_v	u_mass
all-MiniLM-L6-v2	0,5643	-0,7297
all-MiniLM-L12-v2	0,5602	-0,8875
LaBSE	0,4969	-0,6468
all-distilroberta-v1	0,5495	-0,8213
all-mpnet-base-v2	0,5410	-0,9116

Fonte: Próprio autor

Observando-se a Tabela 5 verifica-se que o modelo LaBSE é o único que apresenta o valor do coeficiente *c_v* em que se pode perceber a menor coerência das palavras nos tópicos em relação aos outros modelos de linguagem.

3.2.8.4 Avaliação de desempenho dos modelos de linguagem

A Tabela 6 apresenta o comportamento de todos os modelos de linguagem quando foram submetidos os pedidos de treinamento ou teste. Para cada transformador a tabela apresenta:

- na primeira linha (Original), os resultados logo após o treinamento do modelo de linguagem;
- na segunda linha (Treinamento), os resultados obtidos submetendo aos modelos de linguagem os pedidos de patentes utilizados no treinamento, apenas, para verificar a eficiência dos modelos de linguagem em inferir a estes pedidos que foram utilizados no treinamento os tópicos técnicos criados;

- na terceira linha (Teste), os resultados obtidos submetendo aos modelos de linguagem os pedidos de patentes de teste, sendo estes os percentuais que determinam a eficiência dos modelos.

Tabela 6 – Inferências dos tópicos técnicos gerados pelos modelos

Transformadores	Pedidos submetidos	Acertos Família	Acertos Citados	Acertos Similares
all-MiniLM-L6-v2	Original	86,34%	21,84%	25,68%
	Treinamento	78,64%	21,32%	23,26%
	Teste	72,84%	15,72%	15,79%
all-MiniLM-L12-v2	Original	83,34%	14,54%	18,32%
	Treinamento	78,82%	15,43%	16,53%
	Teste	74,67%	8,55%	11,63%
LaBSE	Original	83,68%	15,30%	23,68%
	Treinamento	74,39%	15,16%	18,75%
	Teste	71,08%	9,86%	12,95%
all-distilroberta-v1	Original	85,40%	21,61%	22,65%
	Treinamento	79,50%	21,19%	20,09%
	Teste	76,64%	14,20%	15,05%
all-mpnet-base-v2	Original	87,92%	21,00%	21,59%
	Treinamento	82,01%	20,20%	19,36%
	Teste	79,26%	13,36%	14,92%

Fonte: Próprio autor

Em todos os modelos de linguagem, tem-se uma pequena queda no desempenho quando comparados os resultados de treinamento (primeira linha) com os resultados submetendo os pedidos de treinamento (segunda linha) aos modelos. Isso indica que não ocorreu *Overfitting*.

A queda do desempenho dos modelos de linguagem com relação a inferência dos pedidos de treinamento para os pedidos de teste ficou entre 2,74% e 7,47%. Ou seja, quando os modelos de linguagem inferem tópicos aos pedidos de teste (terceira linha), eles apresentam um resultado de 2,74% até 7,47% menor.

Os resultados observados estão dentro do esperado: os modelos apresentam um determinado desempenho quando treinados (primeira linha), depois os dados de treinamento são submetidos (segunda linha) e os modelos apresentam queda de desempenho. Posteriormente, para os dados de teste, desconhecidos durante o processo de treinamento, o desempenho cai um pouco mais.

3.2.9 Etapa T9 (escolher modelo de linguagem)

Nesta etapa foram realizadas diversas análises para escolher qual dos modelos de linguagem apresentou melhor desempenho. Também, foram demonstrados alguns dos recursos disponíveis no BERTopic que podem auxiliar no entendimento de novos modelos de linguagem. Cada área do conhecimento irá apresentar as suas próprias particularidades, portanto a diversidade de recursos e a flexibilidade da metodologia precisam ser utilizadas para encontrar o melhor modelo de linguagem possível para uma determinada área do conhecimento.

3.2.9.1 Escolha do modelo de linguagem

A Tabela 7 apresenta os percentuais de acertos dos modelos para os três tipos de pedidos relacionados aos pedidos do conjunto ATInicial com os pedidos de patentes de teste. Nesta tabela, as cores mais escuras indicam um melhor desempenho do modelo. Os modelos de linguagem all-MiniLM-L6-v2, all-distilroberta-v1 e all-mpnet-base-v2 apresentam comportamento semelhante.

Tabela 7 – Comparativo de desempenho com pedidos de teste

Transformadores	Tópicos Técnicos	Acertos Família	Acertos Citados	Acertos Similares
all-MiniLM-L6-v2	273	72,84%	15,72%	15,79%
all-MiniLM-L12-v2	447	74,67%	8,55%	11,63%
LaBSE	168	71,08%	9,86%	12,95%
all-distilroberta-v1	330	76,64%	14,20%	15,05%
all-mpnet-base-v2	384	79,26%	13,36%	14,92%

Fonte: Próprio autor

Na Tabela 8, foram listados os desempenhos dos modelos de linguagem do ponto de vista das famílias dos pedidos. Os pedidos das famílias são os únicos entre família, citados e similares que se pode afirmar que são semelhantes aos pedidos do conjunto ATInicial.

Tabela 8 – Comparativo dos Transformadores x Família

Transformadores	Tópicos Técnicos	Família Original	Família Treinamento	Família Teste
all-MiniLM-L6-v2	273	86,34%	78,64%	72,84%
all-MiniLM-L12-v2	447	83,34%	78,82%	74,67%
LaBSE	168	83,68%	74,39%	71,08%
all-distilroberta-v1	330	85,40%	79,50%	76,64%
all-mpnet-base-v2	384	87,92%	82,01%	79,26%

Fonte: Próprio autor

A Tabela 8 apresenta, na coluna Original, os desempenhos calculados logo após o treinamento dos modelos; na coluna Treinamento, o desempenho resultante da submissão dos dados de treinamento no modelo; e na coluna Teste, o desempenho resultante da submissão dos dados de teste no modelo.

O comportamento dos cinco modelos de linguagem foi semelhante e qualquer um deles poderia ter sido escolhido (ver item 2.1.10). Entretanto, observando as tabelas acima, os seguintes pontos foram ressaltados:

- Na Tabela 6, todos os modelos de linguagem apresentaram queda de desempenho quando foram submetidos aos pedidos utilizados no treinamento (segunda linha), comportamento considerado normal;
- Na Tabela 6, todos os modelos de linguagem apresentaram queda de desempenho quando foram submetidos aos pedidos de teste (terceira linha), comportamento considerado normal;
- Na Tabela 7, os modelos de linguagem com os transformadores all-MiniLM-L12-v2 e LaBSE apresentaram desempenho menor que os outros;
- O modelo de linguagem com o transformador LaBSE apresentou instabilidade durante os testes (ver item 3.2.8.1);
- Na Tabela 7, os três modelos all-MiniLM-L6-v2, all-distilroberta-v1 e all-mpnet-base-v2 apresentam percentuais próximos;
- Na Tabela 8, o modelo de linguagem com o transformador all-mpnet-base-v2 se destacou por apresentar desempenho maior que os outros modelos para as três situações: momento do treinamento, inferência dos pedidos de treinamento e inferência dos pedidos de teste;

- Somente o modelo LaBSE apresentou um desempenho um pouco menor para o *c_v*. Nos outros modelos, tanto o coeficiente de coerência *c_v* como o *u_mass* ficaram muito próximos, não influenciando de forma relevante na escolha do modelo.

Levando em consideração os itens listados acima, optou-se pelo modelo de linguagem com o transformador all-mpnet-base-v2, principalmente, por ele ter apresentado um desempenho ligeiramente melhor que os modelos para os pedidos das famílias de treinamento e teste. Os pedidos das famílias foram considerados mais relevantes, pois são os pedidos que comprovadamente são próximos aos pedidos do conjunto ATInicial.

3.2.9.2 Recursos disponíveis no BERTopic

Muito além da escolha do modelo de linguagem, nesta etapa do módulo de treinamento (MT), o Pesquisador precisa explorar ferramentas disponíveis no modelo de tópicos em uso. Estas ferramentas disponibilizadas pelos desenvolvedores do BERTopic precisam ser analisadas para que se possa realizar melhorias no treinamento dos modelos e disponibilizar recursos úteis ao Usuário do modelo de linguagem. A análise destes recursos será concentrada no modelo de linguagem selecionado all-mpnet-base-v2.

Neste tipo de modelagem, cada tópico é representado por um conjunto de palavras. Nestes exemplos, foram utilizadas 10 palavras para a representação dos tópicos técnicos. Testes com mais e menos palavras foram realizados e as mudanças no desempenho foram irrelevantes, inclusive os desenvolvedores do BERTopic sugerem o uso de 10 palavras.

Análise dos tópicos gerados:

Durante os procedimentos de treinamento, com a submissão de 57.982 pedidos foram gerados 384 tópicos técnicos diferentes no modelo de linguagem com o transformador all-mpnet-base-v2.

Na Figura 29, o gráfico apresenta a distância entre os tópicos e a dimensão dos tópicos gerados no modelo de linguagem. Essa visualização em duas dimensões auxilia para que se possa ter uma visão geral das distâncias entre os tópicos técnicos e por consequência dos documentos que pertencem aos tópicos técnicos.

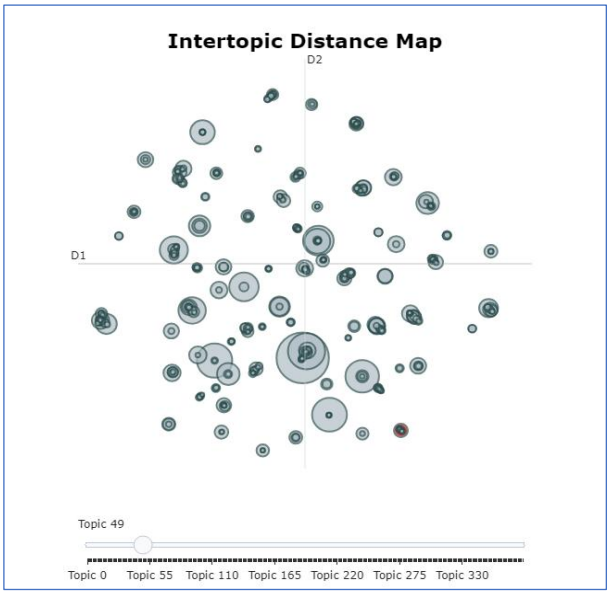


Figura 29 – Mapa de tópicos gerados
Fonte: BERTopic

Cada tópico técnico é associado a um número inteiro apenas para identificação, sendo a representação dos tópicos associada com as dez palavras que representam o tópico.

Para exemplificar, foi realizada uma busca nos tópicos pela palavra *filesystem* que pode ser encontrada nos textos de definição do símbolo de classificação G06F 16/00 da IPC. Os tópicos técnicos relacionados a palavra *filesystem* são: 170, 49, 183, 48 e 175.

A Figura 30 apresenta um zoom de uma região do mapa de tópicos onde se encontra o tópico 49. Os quatro tópicos mais próximos ao tópico 49 são os tópicos 326, 175, 342 e 170.

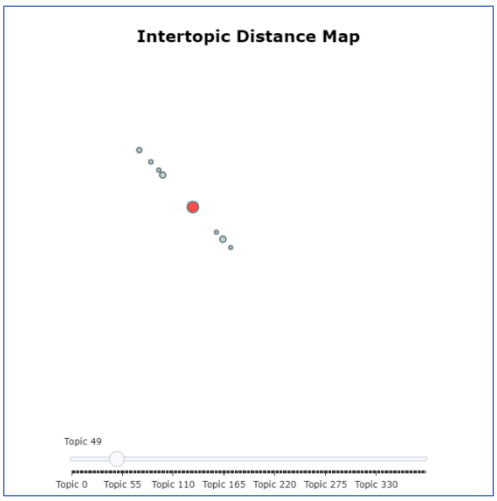


Figura 30 – Zoom do mapa de tópicos
Fonte: BERTopic

As palavras que representam o tópico 49 estão na nuvem de palavras da Figura 31 e pode-se perceber que a palavra pesquisada no exemplo é uma das mais relevantes para o tópico.

Figura 31 – Nuvem de palavras do tópico 49
Fonte: BERTopic

As Figura 32, Figura 33, Figura 34 e Figura 35 apresentam as palavras mais relevantes dos quatro tópicos mais próximos ao tópico 49. Por exemplo, as duas palavras “*file*” e “*files*” podem ser visualizadas em todos os tópicos entre as 10 palavras de suas representações.

Figura 32 – Nuvem de palavras do tópico 175
Fonte: BERTopic

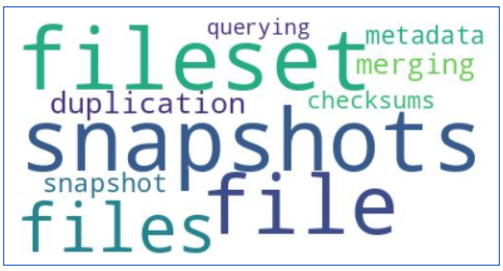


Figura 33 – Nuvem de palavras do tópico 326
Fonte: BERTopic

Figura 34 – Nuvem de palavras do tópico 170
Fonte: BERTopic

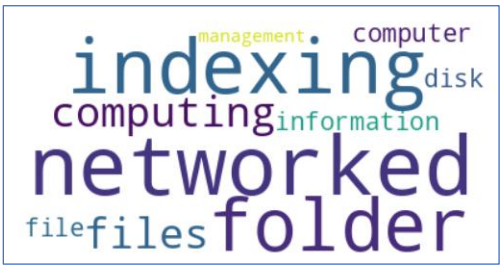


Figura 35 – Nuvem de palavras do tópico 342
Fonte: BERTopic

As nuvens de palavras das figuras acima demostram as dez palavras mais relevantes de cada tópico sendo o tamanho da fonte de cada palavra diretamente proporcional a sua relevância no tópico. Observa-se que os tópicos próximos ao tópico 49 apresentam algumas palavras iguais e outras palavras que os diferenciam.

A nuvem de palavras é uma ferramenta que pode ser utilizada para auxiliar o Pesquisador a identificar novas palavras que podem ser incluídas na lista de *stop words* para que sejam retiradas durante as etapas de treinamento dos modelos de linguagem.

A Figura 36 apresenta a matriz de similaridade entre os tópicos técnicos. Essa matriz apresenta um mapa de calor que relaciona todos os tópicos e agrupa os pares que apresentam similaridades. Com este gráfico o Pesquisador pode analisar quais tópicos apresentam uma maior correlação.

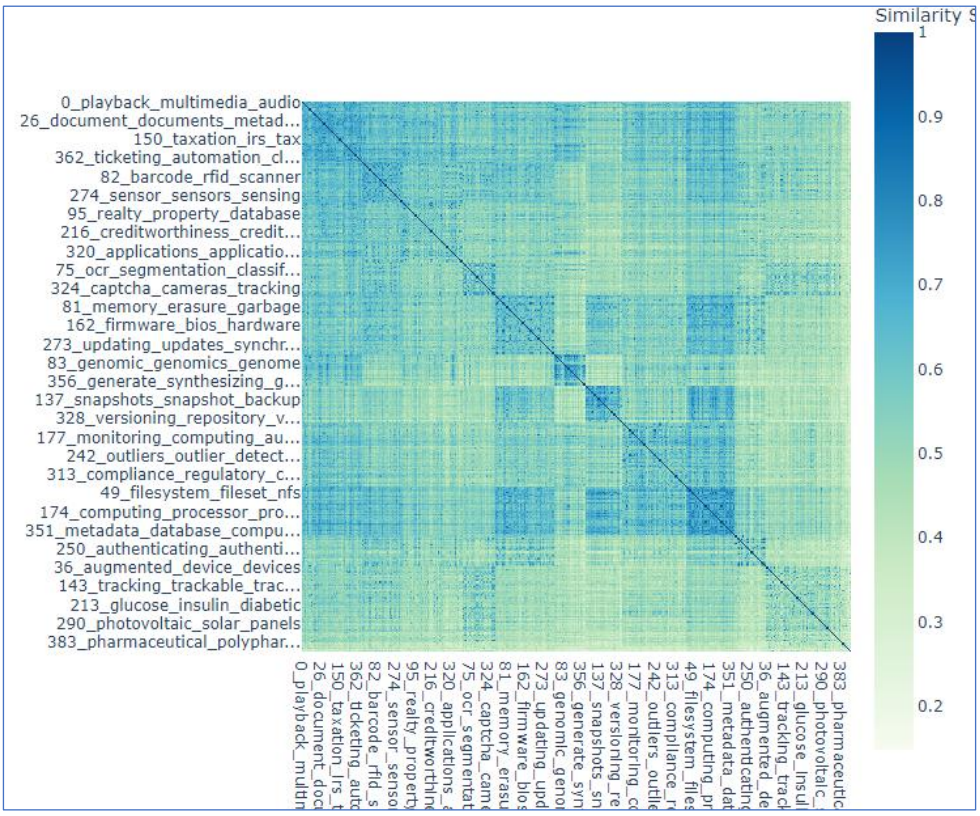


Figura 36 – Matriz de similaridade entre os tópicos técnicos
Fonte: BERTopic

A Figura 37 apresenta parte da hierarquia entre os tópicos técnicos, em que se pode observar a relação entre os tópicos. Observando a hierarquia entre os tópicos o Pesquisador pode verificar quais tópicos apresentam a tendência de formar tópicos maiores.

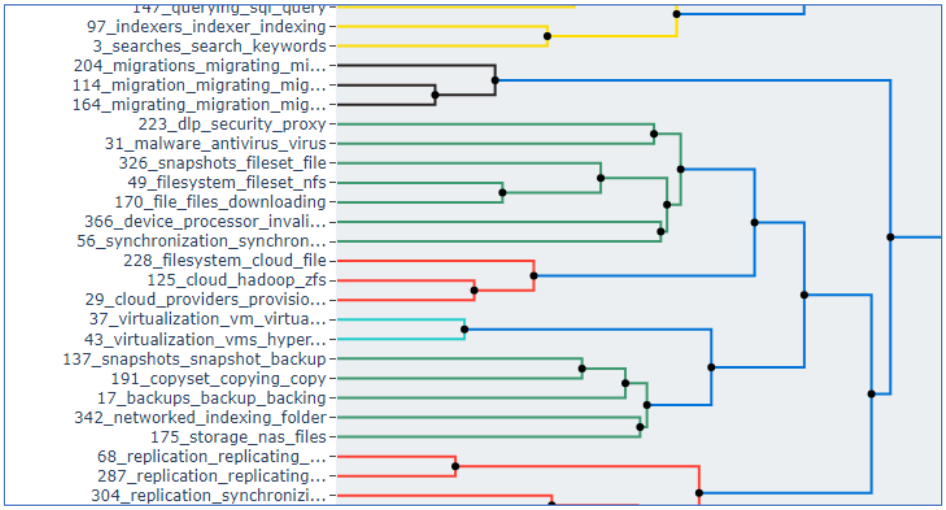


Figura 37 – Hierarquia dos tópicos técnicos
Fonte: BERTopic

A redução de tópicos técnicos é a ferramenta que permite agrupar tópicos próximos e, por consequência, agrupar os documentos que pertencem a estes tópicos técnicos. Essa ferramenta será utilizada a seguir, no módulo de uso (MU) para agrupar os tópicos técnicos na quantidade de tópicos que o Usuário solicitar.

3.3 ETAPAS DE MU – EXEMPLO DETALHADO

As etapas de U1 até U9 foram desenvolvidas obtendo-se, ao final, os pedidos de patentes agrupados em conjuntos de pedidos similares.

Com o objetivo de demonstrar os procedimentos necessários para o módulo de uso do modelo de linguagem, foram realizadas as etapas mantendo a área de conhecimento definida no treinamento do modelo de linguagem.

3.3.1 Etapa U1 (verificar modelos de linguagem disponíveis)

Esta etapa da metodologia utiliza a mesma área definida no treinamento do modelo de linguagem que o Usuário escolher, sendo baseada no símbolo de classificação G06F16/00, bem como utilizou o modelo de linguagem já treinado e escolhido no item 3.2.9.

Neste momento, o Usuário verifica se existe no repositório de modelos de linguagem RAML um modelo de linguagem que seja compatível com os pedidos de patente que o Usuário deseja agrupar. Caso não exista um modelo de linguagem compatível é necessário solicitar que o Pesquisador realize os procedimentos de construção de um novo modelo de linguagem para a área de conhecimento e disponibilize no repositório RAML.

3.3.2 Etapa U2 (selecionar pedidos para agrupar)

A título de demonstração do funcionamento da metodologia, foram selecionados pedidos de patentes pertencentes a mesma área do modelo de linguagem treinado. Este conjunto utilizado no módulo de uso do modelo de linguagem foi denominado de conjunto PCluster.

Os pedidos do conjunto PCluster foram selecionados com as seguintes características:

- Símbolo: G06F 16/00;
- Datas de prioridade: 01/01/2024 até 31/07/2024;
- País de depósito: US (Estados Unidos da América).

A Figura 38 resume as condições utilizadas para a aquisição dos pedidos dos conjuntos PCluster para agrupar. O resultado da pesquisa com os filtros retornou para o conjunto PCluster uma quantidade de 97 pedidos de patentes que precisam ser agrupados. Possivelmente, a quantidade de pedidos encontrada para um período de sete meses foi pequena devido a maioria dos pedidos depositados nesse período ainda não ter sido publicado em 19/08/2024 data em que a pesquisa foi realizada.

Figura 38 – Filtros para PCluster do exemplo detalhado
Fonte: Próprio autor

Em uma situação real, o Usuário precisa reunir os números dos pedidos que ele julga serem compatíveis com o modelo de linguagem selecionado e fornecer para as rotinas essa lista de pedidos. Este Usuário pode ser um chefe de divisão de exame de pedidos de patentes ou mesmo o examinador de patentes que precisa agrupar pedidos objetivando otimizar suas atividades.

3.3.3 Etapa U3 (número de tópicos)

Conforme apresentado anteriormente, a pesquisa realizada em U2 o *GooglePatents* retornou 97 pedidos, os quais precisam ser agrupados. Para dar continuidade ao exemplo detalhado, foi definido que estes pedidos precisam ser distribuídos em 25

agrupamentos (tópicos técnicos). Portanto, o modelo de linguagem com o transformador all-mpnet-base-v2 selecionado no item 3.2.9, que apresentou 384 tópicos técnicos diferentes, foi recalculado e o número de tópicos técnicos passa a ser 25 neste momento. Isso é uma operação realizada no modelo de linguagem já treinado e não apresenta um custo computacional elevado, permitindo ao Usuário testar números de tópicos diferentes se necessário.

Este número de tópicos para agrupamento é uma escolha do Usuário, essa redução no número de tópicos é realizada com base na hierarquia entre os tópicos e é necessária para que uma quantidade pequena de pedidos de patentes seja agrupada. Caso não fosse reduzido o número de tópicos os pedidos de patentes, deste exemplo, poderiam ser inferidos cada um em um tópico diferente.

3.3.4 Etapa U4 (dados dos pedidos)

A Figura 39 apresenta os dados coletados dos pedidos do conjunto PCluster. Este conjunto de dados e metadados coletado precisa ser compatível com os dados dos pedidos utilizados no treinamento do modelo de linguagem para que o modelo possa inferir o pedido em algum tópico técnico. Neste exemplo, somente os textos das reivindicações dos pedidos do conjunto PCluster são essenciais para que possam ser submetidos ao modelo, pois foram só as reivindicações que foram utilizadas no módulo de treinamento (MT) e são estes textos das reivindicações que irão influenciar para que o pedido seja inferido em algum tópico técnico do modelo de linguagem.

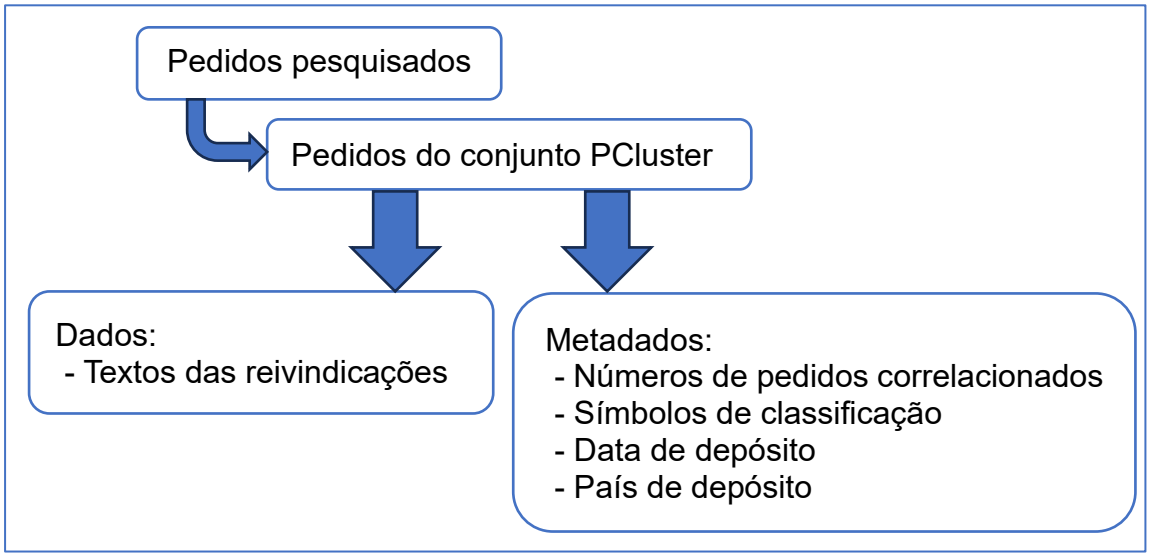


Figura 39 – Pedidos e informações do PCluster coletados
Fonte: Próprio autor

3.3.5 Etapa U5 (armazena dados)

As rotinas armazenam os dados e metadados dos pedidos do conjunto PCluster no banco de dados. A forma de armazenamento utilizada foi muito similar a apresentada na etapa T5 do módulo de treinamento (MT). Para o módulo de uso (MU), são coletados apenas os dados dos pedidos do conjunto PCluster não sendo coletados os dados dos pedidos relacionados como foi realizado em MT. Os pedidos relacionados são utilizados apenas para treinar os modelos de linguagem.

3.3.6 Etapa U6 (prepara dados)

As rotinas precisam realizar os mesmos procedimentos de limpeza nos dados dos pedidos que foram executados para os dados de treinamento. A forma de preparação dos dados foi muito similar à apresentada na etapa T6 do módulo de treinamento (MT).

3.3.7 Etapa U7 (modelo de linguagem)

As rotinas buscam o modelo de linguagem com o transformador all-mpnet-base-v2 do item 3.2.9, definido pelo usuário na etapa U1.

3.3.8 Etapa U8 (submete pedidos ao modelo)

As rotinas submetem os 97 pedidos de patentes ao modelo de linguagem reduzido para 25 tópicos técnicos. O modelo de linguagem infere, a cada pedido de patentes, um tópico técnico, conforme demonstrado na Tabela 9.

O conjunto PCluster foi submetido ao modelo com 97 pedidos e o modelo inferiu tópico em 90 destes pedidos, sendo que 7 pedidos o modelo não atribuiu a nenhum dos tópicos. A técnica de tópicos é baseada na obtenção de escores para que os pedidos de patentes sejam inferidos aos tópicos, possivelmente estes 7 pedidos de patentes não atingiram o escore mínimo necessário para que o modelo de linguagem tivesse condições de agrupar estes documentos.

3.3.9 Etapa U9 (avaliar agrupamentos dos pedidos)

O Usuário precisa verificar se o resultado do agrupamento dos pedidos foi satisfatório. No momento, foram utilizados exemplos para validação da metodologia. Entretanto, para que a etapa de uso da metodologia seja avaliada de forma mais consistente, será necessário um projeto com situações reais, o que não é o escopo deste trabalho.

Semelhante aos recursos para análise dos resultados disponibilizado para o Pesquisador. O Usuário também pode utilizar recursos que podem auxiliar na visualização da proximidade dos pedidos de patentes.

Tabela 9 – Pedidos do conjunto PCluster nos tópicos

Número do tópico	Palavras que representam o tópico	Números dos pedidos atribuídos aos tópicos	Quant.
0	user, content, claim, second, media, method, device, data, plurality, comprising	US20240161149A1, US20240256544A1, US20240256583A1, US20240256595A1, US20240256596A1, US20240256611A1, US20240256629A1, US20240256692A1, US20240257052A1, US20240259446A1, US20240259456A1, US20240259763A1	12
1	data, storage, method, second, claim, file, plurality, comprising, database, set	US20240256384A1, US20240256409A1, US20240256484A1, US20240256487A1, US20240256489A1, US20240256490A1, US20240256493A1, US20240256501A1, US20240256508A1, US20240256521A1, US20240256527A1, US20240256528A1, US20240256537A1, US20240256538A1, US20240256541A1, US20240256568A1, US20240256570A1, US20240256571A1, US20240256576A1, US20240256680A1, US20240257235A1, US20240259209A1, US20240259249A1, US20240259434A1	24
2	vehicle, location, device, information, mobile, claim, method, data, based, second	US20240255291A1, US20240255294A1, US20240256522A1, US20240256579A1, US20240257031A1, US20240257158A1, US20240257288A1, US20240257458A1, US20240259227A1	9
3	search, query, document, text, question, method, language, results, word, claim	US20240256499A1, US20240256506A1, US20240256507A1, US20240256588A1, US20240256592A1, US20240256599A1, US20240256619A1, US20240256620A1, US20240256621A1, US20240256625A1, US20240256628A1, US20240257169A1	12
4	test, code, event, data, method, time, claim, plurality, based, comprising	US20240256566A1, US20240259254A1	2
5	image, display, images, object, reality, touch, user, device, second, information	US20240256101A1, US20240256110A1, US20240257050A1	3
6	model, training, learning, neural, machine, data, feature, food, plurality, network	US11995109B1, US20240256561A1, US20240256598A1, US20240256617A1, US20240256845A1, US20240256920A1, US20240256926A1, US20240256979A1, US20240256982A1, US20240257445A1, US20240257494A1	11
7	medical, patient, healthcare, health, data, clinical, claim, information, surgical, method	US12001464B1, US20240252150A1	2
8	power, energy, estate, building, property, real, hvac, consumption, construction, temperature	US20240257280A1	1
9	canceled, user, device, claim, associated, second, comprising, computer, method, communication	US20240256486A1, US20240256509A1, US20240256514A1, US20240256636A1, US20240256704A1, US20240257266A1, US20240257272A1, US20240259381A1, US20240259419A1, US20240259479A1	10
11	email, domain, message, phishing, mail, address, dns, method, claim, messages	US20240256548A1, US20240257249A1	2
14	robot, rpa, robotic, camera, process, session, automation, object, image, cleaning	US20240253235A1	1
16	jukebox, media, seat, entertainment, players, digital, playlist, recognized, passenger, content	US20240259609A1	1

Fonte: Próprio autor

A Figura 40 apresenta a distância entre os tópicos e a dimensão dos tópicos após a redução de 384 para 25 tópicos técnicos. As dimensões dos tópicos continuam sendo proporcionais à quantidade de pedidos de treinamento classificados em cada tópico.

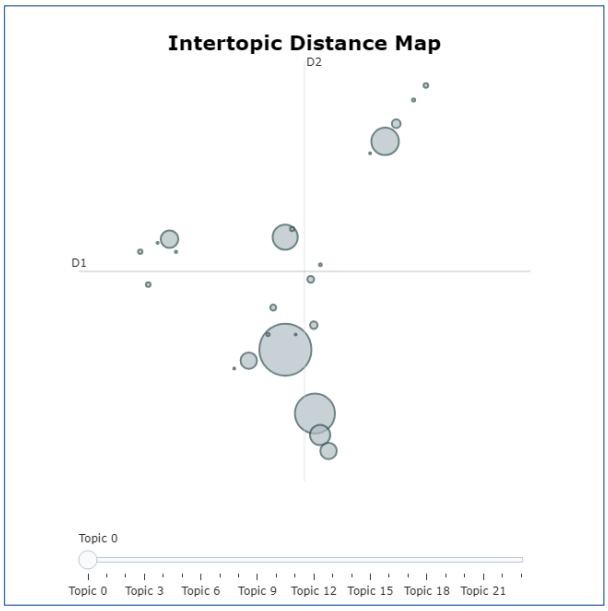


Figura 40 – Mapa reduzido de tópicos gerados
Fonte: BERTopic

A Figura 41 apresenta a nova hierarquia entre os tópicos técnicos com a redução para 25 tópicos. Caso o Usuário entenda que alguns dos pedidos foram inferidos em tópicos isolados o próprio Usuário pode visualizar a hierarquia dos tópicos técnicos e agrupar os pedidos que ficaram isolados com outros pedidos de patentes.

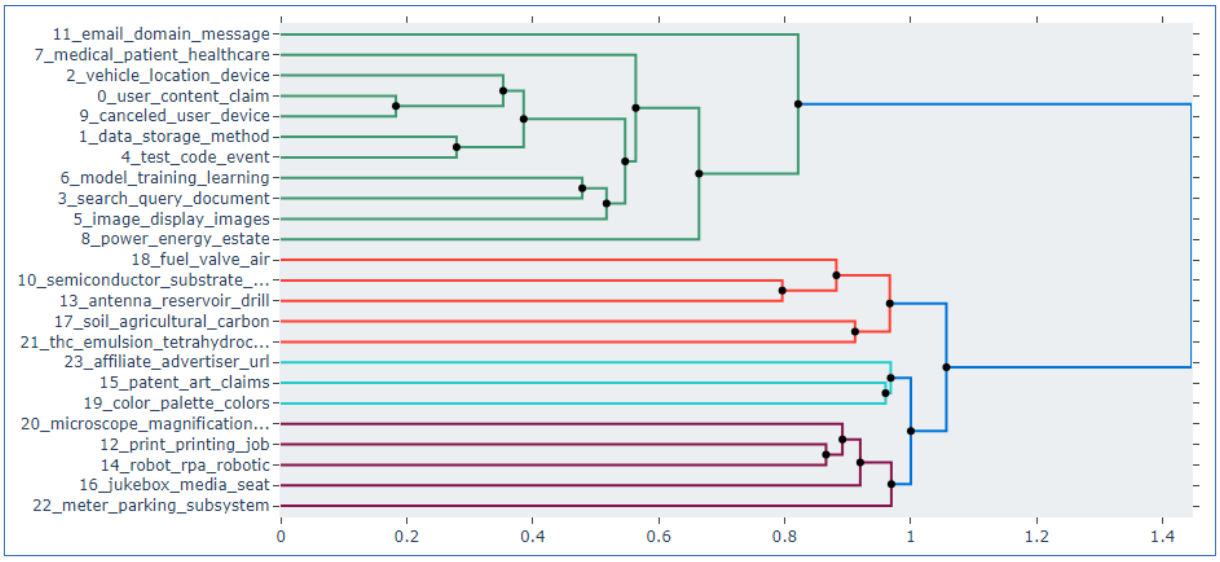


Figura 41 – Hierarquia com 25 tópicos técnicos
Fonte: BERTopic

Semelhante ao realizado no item 3.2.9.2 na etapa de treinamento o Usuário pode fazer uma busca por uma palavra específica, buscando visualizar qual tópico representa uma palavra após a redução de tópicos. A palavra *filesystem* foi novamente pesquisada no modelo de linguagem com 25 tópicos. Retornaram como tópicos mais próximos os tópicos 1, 0, 16 e 4. A Figura 42 apresenta as palavras que representam o tópico 1 que é o mais relevante. Percebe-se que a palavra *filesystem* não aparece entre as 10 palavras principais, mas as palavras da figura são próximas a palavra *filesystem* em termos de área do conhecimento.

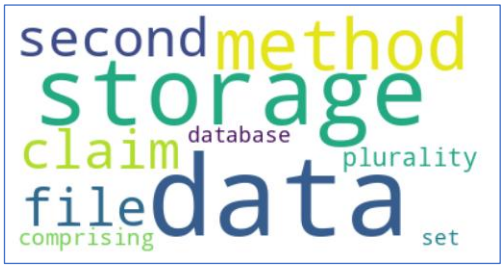


Figura 42 – Nuvem de palavras do tópico 1 com redução
Fonte: BERTopic

Em outro teste, foi realizada a busca nos tópicos por uma segunda palavra *vehicle* e retornaram os tópicos 18, 2 16 e 14 como mais compatíveis com a palavra. A Figura 43 apresenta a nuvem de palavras do tópico 2. O tópico 18 não foi selecionado, pois o modelo não inferiu nenhum dos 97 pedidos para ele.

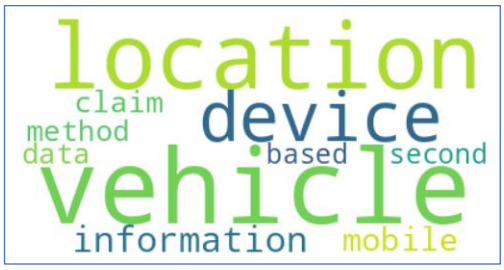


Figura 43 – Nuvem de palavras do tópico 2 com redução
Fonte: BERTopic

A Tabela 10 apresenta os títulos (com tradução automática) dos pedidos que foram agrupados ao tópico 1, o qual foi verificado como o tópico mais compatível com a palavra *filesystem* pesquisada anteriormente. Observou-se que vários dos títulos

apresentam expressões como “banco de dados” ou “armazenamento”, que seriam palavras compatíveis com o termo *filesystem*.

Tabela 10 – Títulos dos pedidos agrupados no tópico 1

Pedido	Título
US20240256384A1	Virtualização em uma rede de armazenamento dispersa utilizando memória local e remota
US20240256409A1	Clustering de servidores em um sistema de computação sob demanda
US20240256484A1	Arquivo de objetos e arquivo baseado em nuvem
US20240256487A1	Atualização de várias cópias de destino criadas a partir de uma única fonte
US20240256489A1	Organização de elementos de dados principais usando uma estrutura de dados em árvore
US20240256490A1	Armazenamento de objetos de dados e servidor para um ambiente de armazenamento em nuvem
US20240256493A1	Rastreamento de arquivos em máquinas clientes sincronizadas com um repositório de sistema de gerenciamento de conteúdo
US20240256501A1	Metadados de dicionário de dados para listagens de mercado
US20240256508A1	Utilização de recursos de banco de dados multilocatário
US20240256521A1	Sistemas e métodos para governança de dados como um serviço
US20240256527A1	Otimização de operações de gravação de banco de dados combinando e paralelizando operações com base em um valor hash de chaves primárias
US20240256528A1	Sistema de banco de dados utilizando indexação probabilística
US20240256537A1	Técnicas para construir linhagens de dados para consultas
US20240256538A1	Filtragem de registros incluídos em objetos de um sistema de armazenamento de objetos com base na aplicação de um pipeline de identificação de registros
US20240256541A1	Execução de consulta por meio de comunicação com um sistema de armazenamento de objetos por meio de um protocolo de comunicação de armazenamento de objetos
US20240256568A1	Aproveitando instantâneos de conjuntos de dados remotos na nuvem
US20240256570A1	Sistemas e métodos de cache
US20240256571A1	Sistemas e métodos de gestão de recursos
US20240256576A1	Sistema de processamento de dados com manipulação de grupos de conjuntos de dados lógicos
US20240256680A1	Propriedade remota e controle de conteúdo de arquivos de mídia em sistemas não confiáveis
US20240257235A1	Sistema e método para agrupar arquivos de dados em grupos relacionados
US20240259209A1	Sistema de blockchain aninhado
US20240259249A1	Suporte a consultas baseadas em graphql em modelos de dados de configuração baseados em yang
US20240259434A1	Prevenção de perda de dados de endpoint de pegada pequena

Fonte: Próprio autor

A Tabela 11 apresenta os títulos dos pedidos agrupados no tópico 2 que é relacionado à pesquisa da palavra *vehicle*. Observando-se os títulos percebe-se expressões como “veículos”, “carros”, “viagem” ou “transporte”.

Tabela 11 – Títulos dos pedidos agrupados no tópico 2

Pedido	Título
US20240255291A1	Mapa esparso para navegação autônoma de veículos
US20240255294A1	Identificação e exibição de caminhos suaves e demarcados
US20240256522A1	Sistemas e métodos para validação autônoma de informações de crowdsourcing e open source
US20240256579A1	Sistema de busca visual para encontrar o destino da viagem
US20240257031A1	Sistemas de elevação
US20240257158A1	Sistemas e métodos de serviços de restauração de carros clássicos
US20240257288A1	Serviço baseado em perfil de usuário de transporte
US20240257458A1	Sistema de realidade cruzada
US20240259227A1	Sistema e método para melhorar a busca de conteúdo selecionando dispositivos de túnel

Fonte: Próprio autor

A comparação entre os pedidos classificados nos tópicos 1 e 2 busca demonstrar que se apresentou uma tendência em agrupar pedidos com assuntos semelhantes. Vale ressaltar que todos os testes realizados têm por objetivo validar a metodologia.

3.4 EXEMPLOS MODIFICADOS

Objetivando realizar um comparativo no desempenho dos modelos com algumas melhorias, foram realizados mais dois testes baseados no transformador de sentenças all-mpnet-base-v2.

No primeiro teste, all-mpnet-base-v2a as modificações foram:

- inclusão de *stop words* que foram detectadas como irrelevantes no exemplo detalhado – essa alteração proporciona a retirada de palavras irrelevantes específicas da área em questão;
- mudança de unigramas para bigramas – essa alteração proporciona que as palavras sejam levadas em consideração conforme o contexto e não mais isoladamente;
- mudança da quantidade mínima de documentos por tópicos técnicos passou de 15 para 25 – essa alteração reduz a quantidade de tópicos técnicos gerados, pois aumenta a exigência de quantidade de pedidos em um tópico durante o treinamento.

No segundo teste, all-mpnet-base-v2b as modificações foram:

- mesmas modificações inseridas no primeiro teste com modificações;
- inclusão da lematização das palavras – essa alteração reduz a quantidade de variações de palavras que possivelmente tem um significado muito semelhante.

A lista de stop words adicionais foi obtida apenas pela observação de palavras que aparentemente não eram relevantes, mas se apresentaram em grande quantidade em mapas de palavras de alguns tópicos técnicos. As palavras foram: *application*, *method*, *claim*, *data*, *canceled*, *wherein*, *one*, *second*, *first*, *system*, *information*, *least*.

A Tabela 12 apresenta as coerências *c_v* e *u_mass* do modelo baseado no transformador de sentenças all-mpnet-base-v2 e suas variações all-mpnet-base-v2a e all-mpnet-base-v2b. O teste all-mpnet-base-v2a apresentou uma pequena melhora de desempenho em *c_v* com relação aos outros. Neste teste, foram incluídas algumas *stop words*, bigramas e mudança do mínimo de documentos por tópico. As modificações não geraram melhorias do ponto de vista da coerência *u_mass*.

Tabela 12 – Medidas de coerência das variações

Transformadores	c_v	u_mass
all-mpnet-base-v2	0,5410	-0,9116
all-mpnet-base-v2a	0,5578	-0,7672
all-mpnet-base-v2b	0,5376	-0,7744

Fonte: Próprio autor

Na Tabela 13 se pode verificar que as variações atribuídas ao modelo de linguagem all-mpnet-base-v2 apresentaram uma queda de desempenho com relação aos pedidos de família, mas apresentaram melhora de desempenho para citados e similares. A melhoria no agrupamento destes tipos de pedidos relacionados pode ser mais relevante do que melhorias na proximidade de pedidos das famílias.

Tabela 13 – Inferências dos tópicos técnicos gerados pelos modelos

Transformadores	Pedidos submetidos	Acertos Família	Acertos Citados	Acertos Similares
all-mpnet-base-v2	Original	87,92%	21,00%	21,59%
	Treinamento	82,01%	20,20%	19,36%
	Teste	79,26%	13,36%	14,92%
all-mpnet-base-v2a	Original	88,37%	25,19%	24,69%
	Treinamento	81,12%	24,29%	22,48%
	Teste	78,14%	17,62%	17,18%
all-mpnet-base-v2b	Original	88,31%	24,50%	23,78%
	Treinamento	77,73%	23,27%	21,64%
	Teste	77,14%	15,25%	15,35%

Fonte: Próprio autor

A Tabela 14 apresenta o comparativo dos resultados do modelo escolhido e de suas variações modificadas. Pode-se perceber que, com as modificações, houve mudanças nos percentuais de acerto. Houve uma melhora no reconhecimento da família no momento do treinamento, mas houve uma pequena queda no desempenho quando os pedidos de treinamento e teste são inferidos aos modelos.

Tabela 14 – Comparativo dos Transformadores com modificações

Transformadores	Tópicos Técnicos	Família Original	Família Treinamento	Família Teste
all-mpnet-base-v2	384	87,92%	82,01%	79,26%
all-mpnet-base-v2a	235	88,37%	81,12%	78,14%
all-mpnet-base-v2b	240	88,31%	77,73%	77,14%

Fonte: Próprio autor

A realização destes testes adicionais com modificações, visou demonstrar a versatilidade da metodologia proposta em permitir modificações e realizar comparativos entre os modelos de linguagem construídos e suas variações.

Uma quantidade muito maior de testes com as variações possíveis dos modelos de linguagem precisa ser realizada para que se possa concluir quais modificações realmente apresentam melhora de desempenho dos modelos.

CONCLUSÃO

O objetivo deste trabalho foi propor uma metodologia para construir modelos de linguagem que agrupem pedidos em tópicos técnicos específicos, buscando gerar subsídios para a distribuição dos pedidos para exame, por meio dos conteúdos técnicos dos referidos pedidos de patentes, tais como: reivindicações, relatório descritivo e resumo.

A principal contribuição da presente tese está no desenvolvimento e validação da metodologia de agrupamento de pedidos de patentes em tópicos. A metodologia foi dividida em dois módulos. O módulo de treinamento que precisa ser desenvolvido por um cientista de dados (Pesquisador) e o módulo de uso que foi estruturado para o chefe de divisão ou examinador de patentes (Usuário). Cada um dos dois módulos da metodologia foi subdividido em nove etapas distintas buscando facilitar o entendimento dos procedimentos. A estrutura de integração dos dois módulos conta com um repositório por área de conhecimento de modelos de linguagem (RAML).

Dentre os resultados considerados relevantes, pode-se listar os seguintes itens:

- com a escolha dos pedidos da família como balizadores do desempenho dos modelos, obteve-se um desempenho entre 83% e 87%, conforme o modelo, no agrupamento dos pedidos do conjunto ATInicial com seus pedidos de família no momento do treinamento e um desempenho entre 71% e 79% com pedidos de teste;
- o desempenho do agrupamento dos pedidos ATInicial com seus pedidos citados e similares foi semelhante, entre 14% e 25%. Esse resultado foi considerado relevante por demonstrar a sensibilidade dos modelos;
- devido ao desempenho maior no agrupamento para os pedidos da família e menor para os citados e similares, considera-se que os modelos apresentaram sensibilidade aos dados, não ficando sempre com mesmo desempenho para qualquer pedido;
- observando-se o comportamento das curvas de erro RMSE, verifica-se que o aumento gradativo do número de pedidos de treinamento resulta em um erro RMSE menor. Presume-se que aumentando ainda mais o número de pedidos de treinamento a tendência de queda do erro continue;

- comparando-se os cinco modelos de linguagem percebe-se que o modelo com transformador de sentenças LaBSE apresentou a mais baixa medida de coerência *c_v* e o pior desempenho no agrupamento dos pedidos do conjunto ATInicial com os pedidos das suas famílias, sendo assim, esse modelo foi descartado;
- ao realizar a inferência dos dados de treinamento e dos dados de teste aos modelos treinados, em todos os casos, ocorreu uma queda de desempenho entre 2,74% e 7,47%. Pode-se deduzir que não ocorreu *overfitting*; e
- o modelo escolhido obteve o melhor desempenho com relação a família dos pedidos nas três situações, com 87% de acertos ao final do treinamento, 82% com os dados de treinamento e 79% com os dados de teste.

Observando-se os resultados obtidos, foram realizadas as seguintes argumentações:

- o uso de uma base de dados consolidada facilitou o desenvolvimento por ter os pedidos do conjunto ATInicial e seus relacionados;
- a comprovação de que o aumento de pedidos de treinamento melhora o desempenho dos modelos;
- o uso da família como principal parâmetro, para confirmar que o modelo conseguiu agrupar pedidos do ATInicial com os pedidos relacionados;
- o uso dos citados e similares para demonstrar que os modelos são sensíveis aos textos diferentes, não ficando similares a todo tipo de texto;
- o cálculo do RMSE demostrou o desempenho dos modelos;
- as medidas de coerência foram comprovadas quando um dos modelos apresentou um valor menor que os outros e esse mesmo modelo teve o pior desempenho; e
- a opção pela separação entre um módulo de treinamento, voltado para o Pesquisador e um módulo de uso voltado para o Usuário, foi utilizada para demonstrar com clareza que os dois módulos apresentam semelhanças em sua estrutura. Sendo, o módulo de treinamento executado eventualmente e o módulo de uso utilizado frequentemente para agrupar novos documentos.

A modularidade da técnica de tópicos BERTopic proposta por Grootendorst (2022) foi fundamental para a realização deste trabalho. Isso porque facilitou a realização dos

treinamentos e testes, possibilitando a perspectiva de trabalhos futuros, bem como justificando a própria estrutura de módulos e etapas proposta na metodologia.

A validação dos modelos de linguagem foi baseada na capacidade do modelo em agrupar os pedidos originalmente pesquisados (ATInicial) com seus respectivos pedidos das famílias, citados e similares. Observou-se que, para todos os modelos, o percentual de acerto no agrupamento de pedidos do conjunto ATInicial com pedidos de suas famílias foi muito maior do que o agrupamentos de pedidos do conjunto ATInicial com pedidos citados e similares. Esses resultados demonstram a sensibilidade dos modelos de linguagem, acertando com pedidos das famílias, que certamente, são próximos dos pedidos do conjunto ATInicial, bem como acertando menos com pedidos citados e similares. Em outras palavras, os pedidos de patentes relacionados com as famílias foram utilizados como balizadores do desempenho dos modelos de linguagem no agrupamento dos pedidos. Um alto índice de agrupamento dos pedidos do conjunto ATInicial com pedidos das suas famílias indica que os agrupamentos foram eficientes.

Os pedidos citados e similares que foram relacionados aos pedidos do conjunto ATInicial durante o andamento dos processos dos pedidos possuem um relacionamento muito mais subjetivo que o relacionamento dos pedidos das famílias e, por isso, é razoável serem mais distantes.

Em Whalen *et al.* (2020), os autores argumentam que a quantidade de documentos disponíveis e seus metadados tem aumentado muito nos últimos anos. Em Harispe *et al.* (2015), os autores entendem que a noção de semelhança só faz sentido quando considerada dentro de uma representação parcial. Em Thakur *et al.* (2020), os autores afirmam que a utilização de dados aleatórios para o treinamento não geram um bom desempenho. Este trabalho, confirmou essas afirmativas quando utilizou metadados como condições de contorno, coletando informações de forma aleatória, mas com delimitações. Por exemplo, os símbolos de classificação como o ponto de partida para selecionar os pedidos.

Em Whalen *et al.* (2020), os autores utilizam uma grande quantidade de pedidos de patentes para avaliar as similaridades dos documentos, demonstrando que um número maior de documentos melhora o desempenho do sistema em encontrar documentos similares. No presente trabalho, em todos os modelos de linguagem

construídos e testados, observou-se que com o aumento no número de amostras o desempenho do modelo em agrupar pedidos do ATInicial com seus relacionados aumentou, confirmando a necessidade de utilizar um número elevado de documentos nos treinamentos dos modelos de linguagem.

Os modelos de linguagem construídos no BERTopic apresentaram tendências de queda nos erros com o aumento no número de amostras de treinamento. Isso indica que um aumento no número de amostras de treinamento pode gerar modelos com melhor desempenho que os apresentados.

Analisando as medidas de coerência *c_v* dos modelos treinados, pode-se verificar que o modelo de linguagem com o transformador de sentença LaBSE apresenta o menor valor em relação aos outros modelos de linguagem (ver Tabela 5) e o mesmo modelo apresentou o pior desempenho no agrupamento dos pedidos do conjunto ATInicial com os pedidos relacionados (ver Tabela 14). Confirmou-se a relevância das medidas de coerência em avaliar os modelos de linguagem apresentada em Mifrah (2020), Newman *et al.* (2010) e Rahimi *et al.* (2024).

Em Younge; Kuhn (2016), os autores alegam que a medida de similaridades baseada em textos funciona melhor que a classificação dos pedidos. O presente trabalho utilizou pedidos que foram classificados em G06F16/00 e pela modelagem de tópicos o modelo de linguagem selecionado subdividiu-se em 384 tópicos diferentes. Entende-se que a afirmativa dos autores foi comprovada na medida que um grupo principal da IPC tem subdivisões, mas o modelo de tópicos com 384 tópicos reagrupou os pedidos de patentes levando em consideração a similaridade semântica.

Em Hain *et al.* (2022), foram utilizados os textos dos resumos para verificar a similaridade das tecnologias dos pedidos de patentes e foi sugerido que os textos das reivindicações seriam adequados para encontrar infrações. Entende-se que os resultados obtidos, na presente pesquisa, com o uso das reivindicações vão no mesmo sentido dos referidos autores, por terem sido confirmadas as similaridades dos pedidos com o uso das reivindicações.

Limitações:

A principal limitação da metodologia é a necessidade de realizar treinamentos de forma isolada para cada área do conhecimento. Portanto, ainda precisam ser analisados quais seriam os melhores símbolos de classificação para construir os modelos que contemplem todas as áreas.

Obviamente, a capacidade de processamento é sempre um limitante, quando se trata de metodologias que exigem grandes quantidades de dados.

Propostas de trabalhos futuros:

Utilizando-se a metodologia proposta, sugere-se realizar testes de desempenho dos modelos alterando as características de cada um de seus módulos. Por exemplo, gerando modelos com conjuntos de palavras de parada (*stop words*) específicos para cada área do conhecimento. Tal sugestão baseia-se em Sarica e Luo (2021), na qual os autores demonstraram que a retirada de palavras de parada específicas de uma área do conhecimento geram aumento de desempenho.

Conforme abordado em Arts, Hou e Gomez, (2021), a lematização e a estemização são recursos que tendem a reduzir o número de palavras diferentes em um texto. O uso destes recursos pode tornar a etapa de representação dos tópicos mais eficiente, considerando que as palavras que irão representar cada tópico serão mais abrangentes. Por isso, sugere-se na realização de novos trabalhos que os recursos da lematização e da estemização sejam utilizados com maior frequência em um maior número de testes, a fim de comprovar a teoria dos referidos autores.

A utilização da técnica *n-grama* pode tornar o treinamento mais pesado computacionalmente. Entretanto, parece ser bastante promissora, pois faz com que o modelo de linguagem realize uma avaliação não apenas de palavras isoladas, mas de palavras em um contexto levando em consideração *n* palavras. Dessa forma, sugere-se a realização de estudos que utilizem a técnica *n-grama*, a fim de comprovar a eficiência da referida técnica, bem como formas de utilizá-la deixando o treinamento mais leve computacionalmente.

Outra abordagem que pode ser desenvolvida é a representação de múltiplos tópicos. Nesta abordagem, um documento pode ser inferido a mais de um tópico o que

aparenta ser muito relevante. Por exemplo, uma reivindicação de uma patente pode tratar de diversos assuntos e poderia ser, dependendo do ponto de vista, inferida em tópicos distintos. Por isso, sugere-se estudos que utilizem de forma mais aprofundada a representação de múltiplos tópicos, suportada pelo BERTopic.

A real validação do módulo de uso da metodologia só será possível com o desenvolvimento de um projeto com situações reais, nas quais os usuários possam avaliar se os documentos foram agrupados de forma coerente.

Outros testes com as variações dos modelos de linguagem precisam ser realizados, para que se possa concluir quais modificações realmente apresentam efeitos na melhora de desempenho dos modelos.

REFERÊNCIAS

ABRANTES, Paula Cotrim de *et al.* Estudo comparativo de bases gratuitas de patentes: patentscope (wipo), espacenet (epo), buscaweb (inpi/br). **Revista Tecnologia e Sociedade**, [*s. l.*], v. 18, n. 54, p. 125–142, 2022. Disponível em: https://doi.org/10.3895/RTS.V18N54.14516. Acesso em: 7 abr. 2023.

ACKERMAN, Christopher M. Representation Tuning. [*s. l.*], 2024. Disponível em: https://arxiv.org/abs/2409.06927v1. Acesso em: 6 out. 2024.

AHARONSON, Barak S.; SCHILLING, Melissa A. Mapping the technological landscape: Measuring technology distance, technological footprints, and technology evolution. **Research Policy**, [*s. l.*], v. 45, n. 1, p. 81–96, 2016. Disponível em: https://doi.org/10.1016/j.respol.2015.08.001

AKKURT, Furkan *et al.* Evaluating the Quality of a Corpus Annotation Scheme Using Pretrained Language Models. *In*: , 2024. **2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC-COLING 2024 - Main Conference Proceedings**. [*S. l.*]: European Language Resources Association (ELRA), 2024. p. 6504–6514.

ARTS, Sam; HOU, Jianan; GOMEZ, Juan Carlos. Natural language processing to identify the creation and impact of new technologies in patent text: Code, data, and new measures. **Research Policy**, [*s. l.*], v. 50, n. 2, p. 104144, 2021. Disponível em: https://doi.org/10.1016/J.RESPOL.2020.104144

BARBOSA, Denis Borges. Uma introdução à propriedade intelectual. **Lumen Juris**, [*s. l.*], p. 1–951, 2010.

BENNER, Mary; WALDFOGEL, Joel. Close to you? Bias and precision in patent-based measures of technological proximity. **Research Policy**, [*s. l.*], v. 37, n. 9, p. 1556–1567, 2008. Disponível em: https://doi.org/10.1016/j.respol.2008.05.011

BOON, Byungun; PARK, Yongtae. A systematic approach for identifying technology opportunities: Keyword-based morphology analysis. **Technological Forecasting and Social Change**, [*s. l.*], v. 72, n. 2, p. 145–160, 2005. Disponível em: https://doi.org/10.1016/j.techfore.2004.08.011

BRASIL. **Lei 9.279 de 14 de maio de 1996**. Regula direitos e obrigações relativos à propriedade industrial. Brasília: Presidência da República, 1996. Disponível em: https://www.planalto.gov.br/ccivil_03/Leis/L9279.htm. Acesso em: 21 abr. 2023.

CHOLLET, François. **Deep Learning with Python, Second Edition**. [*S. l.: s. n.*], 2021. *E-book*.

COOKE, Neil Edward John; GILLAM, Lee. **System , process and method for the detection of common content in multiple documents in an electronic system**. US9760548B2. Concessão: 2017.

DEVLIN, Jacob *et al.* BERT: Pre-training of deep bidirectional transformers for language understanding. **NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference**, [*s. l.*], v. 1, n. Mlm, p. 4171–4186, 2019.

ECKARDT, Ralph W. *et al.* **Method and apparatus for selecting, analyzing, and visualizing related database records asa network**. US7672950B2. Concessão: 2010.

ECKARDT, Ralph W. *et al.* **Methods of providing network graphical representation of**

database records. US10878016B2. Concessão: 2020.

FILHO, Mario. **RMSE (Raiz do Erro Quadrático Médio) em Machine Learning**. [*S. l.*], 2023. Disponível em: https://mariofilho.com/rmse-raiz-do-erro-quadratico-medio-em-machine-learning/. Acesso em: 7 out. 2024.

GERKEN, Jan M.; MOEHRLE, Martin G. A new instrument for technology monitoring: Novelty in patents measured by semantic patent analysis. **Scientometrics**, [*s. l.*], v. 91, n. 3, p. 645–670, 2012. Disponível em: https://doi.org/10.1007/s11192-012-0635-7

GOOD, Jeff. **A gentle introduction to metadata**. [*S. l.*], 2003. Disponível em: https://www.researchgate.net/publication/283645182_A_gentle_introduction_to_metadata. Acesso em: 7 out. 2024.

GRANATYR, Jones. **Deep Learning com Python de A a Z - O Curso Completo**. [*S. l.*], [*s. d.*]. Disponível em: https://iaexpert.academy/cursos-online-assinatura/deep-learning-python-az-curso-completo/?doing_wp_cron=1686333335.5114710330963134765625. Acesso em: 9 jun. 2023.

GROOTENDORST, Maarten. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. [*s. l.*], 2022. Disponível em: http://arxiv.org/abs/2203.05794

GUAN, Jian Cheng; YAN, Yan. Technological proximity and recombinative innovation in the alternative energy field. **Research Policy**, [*s. l.*], v. 45, n. 7, p. 1460–1473, 2016. Disponível em: https://doi.org/10.1016/j.respol.2016.05.002

HAIN, Daniel S. *et al.* A text-embedding-based approach to measuring patent-to-patent technological similarity. **Technological Forecasting and Social Change**, [*s. l.*], v. 177, n. February, p. 121559, 2022. Disponível em: https://doi.org/10.1016/j.techfore.2022.121559

HARISPE, Sébastien *et al.* **Semantic Similarity from Natural Language and Ontology Analysis**. [*S. l.: s. n.*], 2015. ISSN 19474040.v. 8 Disponível em: https://doi.org/10.2200/S00639ED1V01Y201504HLT027

HSU, Po-Hsuan *et al.* Deep Learning, Text, and Patent Valuation. **SSRN Electronic Journal**, [*s. l.*], 2021. Disponível em: https://doi.org/10.2139/ssrn.3758388

INPI. Manual Básico para Proteção por Patentes de Invenções , Modelos de Utilidade e Certificados de Adição. **INPI**, [*s. l.*], v. 2021, 2021.

JACKSI, Karwan; SALIH, Niyaz. State of the art document clustering algorithms based on semantic similarity. **Jurnal Informatika**, [*s. l.*], v. 14, n. 2, p. 58, 2020. Disponível em: https://doi.org/10.26555/jifo.v14i2.a17513

JEONG, Seongkyoon; LEE, Sungki. What drives technology convergence? Exploring the influence of technological and resource allocation contexts. **Journal of Engineering and Technology Management - JET-M**, [*s. l.*], v. 36, p. 78–96, 2015. Disponível em: https://doi.org/10.1016/j.jengtecman.2015.05.004

JOSÉ, Maria *et al.* Processamento de linguagem natural, linguística de corpus e estudos linguísticos : uma parceria bem-sucedida. [*s. l.*], 2015. Disponível em: https://lume.ufrgs.br/handle/10183/169398. Acesso em: 2 out. 2024.

KNIGHT, Michelle. **Fundamentals of Metadata Management - DATAVERSITY**. [*S. l.*], 2023. Disponível em: https://www.dataversity.net/fundamentals-metadata-management/. Acesso em: 19 abr. 2023.

LEE, Won Sang; HAN, Eun Jin; SOHN, So Young. Predicting the pattern of technology convergence using big-data technology on large-scale triadic patents. **Technological Forecasting and Social Change**, [*s. l.*], v. 100, p. 317–329, 2015. Disponível em:

https://doi.org/10.1016/j.techfore.2015.07.022

LIMA, Dahisy; MIRANDA, Daniel. **Introdução à Topologia Algébrica**. [*S. l.*], 2023. Disponível em: https://danielmiranda.prof.ufabc.edu.br/topologia-algebrica/topologia-algebrica.pdf. Acesso em: 9 out. 2024.

MASTROGIORGIO, Mariano; GILSING, Victor. Innovation through exaptation and its determinants: The role of technological complexity, analogy making & patent scope. **Research Policy**, [*s. l.*], v. 45, n. 7, p. 1419–1435, 2016. Disponível em: https://doi.org/10.1016/j.respol.2016.04.003

MCINNES, Leland; HEALY, John; MELVILLE, James. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. [*s. l.*], 2020. Disponível em: https://arxiv.org/abs/1802.03426v3. Acesso em: 2 out. 2024.

MIFRAH, Sara. Topic Modeling Coherence: A Comparative Study between LDA and NMF Models using COVID'19 Corpus. **International Journal of Advanced Trends in Computer Science and Engineering**, [*s. l.*], v. 9, n. 4, p. 5756–5761, 2020. Disponível em: https://doi.org/10.30534/ijatcse/2020/231942020

MOREIRA JR., Cesar Vianna. **Busca pelo equilíbrio da carga de exame : Uma metodologia para avaliação e distribuição de Pedidos de Patente**. 188 f. 2021. [*s. l.*], 2021. Disponível em: https://www.gov.br/inpi/pt-br/servicos/a-academia/arquivo/teses/tese_cesar-vianna-moreira-junior.pdf

NAGHSHZAN, AmirHossein; RATTE, Sylvie. Enhancing API Documentation through BERTopic Modeling and Summarization. **arXiv**, [*s. l.*], v. 2308, 2023. Disponível em: http://arxiv.org/abs/2308.09070

NEWMAN, David *et al.* Automatic evaluation of topic coherence. **NAACL HLT 2010 - Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Proceedings of the Main Conference**, [*s. l.*], n. June, p. 100–108, 2010.

OLUMIDE, Shittu. **Root Mean Square Error (RMSE) In AI: What You Need To Know**. [*S. l.*], 2023. Disponível em: https://arize.com/blog-course/root-mean-square-error-rmse-what-you-need-to-know/.

PREMEBIDA, Sthefanie Monica. **Guia de NLP - conceitos e técnicas**. [*S. l.*], 2021. Disponível em: https://www.alura.com.br/artigos/guia-nlp-conceitos-tecnicas. Acesso em: 2 dez. 2021.

RAHIMI, Hamed *et al.* Contextualized Topic Coherence Metrics. **EACL 2024 - 18th Conference of the European Chapter of the Association for Computational Linguistics, Findings of EACL 2024**, [*s. l.*], p. 1760–1773, 2024.

SARICA, Serhad; LUO, Jianxi. Stopwords in technical language processing. **PLOS ONE**, [*s. l.*], v. 16, n. 8, p. e0254937, 2021. Disponível em: https://doi.org/10.1371/JOURNAL.PONE.0254937. Acesso em: 5 out. 2024.

THAKUR, Nandan *et al.* Augmented SBERT: Data Augmentation Method for Improving Bi-Encoders for Pairwise Sentence Scoring Tasks. **NAACL-HLT 2021 - 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference**, [*s. l.*], p. 296–310, 2020. Disponível em: https://doi.org/10.18653/v1/2021.naacl-main.28. Acesso em: 1 out. 2024.

WALSH, John P.; LEE, You Na; JUNG, Taehyun. Win, lose or draw? the fate of patented inventions. **Research Policy**, [*s. l.*], v. 45, n. 7, p. 1362–1373, 2016. Disponível em: https://doi.org/10.1016/j.respol.2016.03.020

WHALEN, Ryan *et al.* Patent Similarity Data and Innovation Metrics. **Journal of Empirical Legal Studies**, [*s. l.*], v. 17, n. 3, p. 615–639, 2020. Disponível em: https://doi.org/10.1111/jels.12261

WIPO. **Guide to the International Patent Classification**. [*S. l.*]: Geneva, Switzerland : World Intellectual Property Organization, 2023., 2023. Disponível em: https://doi.org/10.34667/tind.48084

WIPO. **Standards – ST.9**. Recommendation concerning bibliographic data on and relating to patents and SPCS. **WIPO**, 2013. Seção June, Disponível em: https://www.wipo.int/export/sites/www/standards/en/pdf/03-09-01.pdf

WIPO. **WIPO Patent Drafting Manual**. 2nd. ed. [*S. l.*]: [Geneva, Switzerland] : [World Intellectual Property Organization (WIPO)], 2022., 2022. Disponível em: https://doi.org/10.34667/tind.44657

YOON, Janghyeok; KIM, Kwangsoo. Detecting signals of new technological opportunities using semantic patent analysis and outlier detection. **Scientometrics**, [*s. l.*], v. 90, n. 2, p. 445–461, 2012. Disponível em: https://doi.org/10.1007/s11192-011-0543-2

YOON, Janghyeok; PARK, Hyunseok; KIM, Kwangsoo. Identifying technological competition trends for R&D planning using dynamic patent maps: SAO-based content analysis. **Scientometrics**, [*s. l.*], v. 94, n. 1, p. 313–331, 2013. Disponível em: https://doi.org/10.1007/s11192-012-0830-6

YOUNGE, Kenneth A.; KUHN, Jeffrey M. Patent-to-Patent Similarity: A Vector Space Model. **SSRN Electronic Journal**, [*s. l.*], 2016. Disponível em: https://doi.org/10.2139/ssrn.2709238

Apêndice A – Classificação IPC

Objetivos da IPC:

A IPC tem como principal objetivo o estabelecimento de uma ferramenta de busca eficaz para a recuperação de documentos de patentes (WIPO, 2023).

A IPC ainda busca servir como:

- um instrumento para o arranjo ordenado de documentos de patente a fim de facilitar o acesso às informações tecnológicas e legais contidas nos mesmos;

- uma base para a disseminação seletiva de informações a todos os usuários das informações de patentes;

- uma base para investigar o estado da técnica em determinados campos da tecnologia;

- uma base para a elaboração de estatísticas sobre propriedade industrial que permitam a avaliação do desenvolvimento tecnológico em diversas áreas (WIPO, 2023).

Breve histórico da IPC:

O texto da primeira edição da IPC teve início na Convenção Europeia sobre Classificação em 1954. Essa primeira edição esteve em vigor de 1968 até 1974. Posteriormente, foram criadas mais cinco edições até a grande reforma da IPC de 2006 (WIPO, 2023).

Após a reforma da IPC, em 2006, esta foi dividida em níveis básico e avançado de 2006 até 2010. Já em 2011 a abordagem de níveis foi descontinuada e as novas versões começaram a ser publicadas com indicação de ano e mês, por exemplo, a versão de 2023 é citada como IPC 2023.01 (WIPO, 2023).

Estrutura da IPC:

Os primeiros níveis da hierarquia da IPC são constituídos por um símbolo que facilita a indexação e um título que descreve sobre o item. Por exemplo, a seção A, apresenta o símbolo “A” e o título “NECESSIDADES HUMANAS”.

Seção:

A classificação está dividida em oito seções. Estas seções são o nível mais alto da hierarquia de classificação e abarcam todos os conjuntos de conhecimentos. Sendo assim, todos os pedidos se encaixam em alguma destas seções (WIPO, 2023). Seguem as seções da IPC:

A - NECESSIDADES HUMANAS
B - OPERAÇÕES DE PROCESSAMENTO; TRANSPORTE
C - QUÍMICA; METALURGIA
D - TÊXTEIS; PAPEL
E - CONSTRUÇÕES FIXAS
F - ENGENHARIA MECÂNICA; ILUMINAÇÃO; AQUECIMENTO; ARMAS; EXPLOSÃO
G - FÍSICA
H - ELETRICIDADE

Classe:

Cada uma das seções é dividida em classes, as quais são o segundo nível na hierarquia da classificação. As classes são constituídas pelo símbolo da classe que consiste em dois dígitos numéricos (por exemplo, H01) e título da classe (por exemplo, H01 – Elementos Elétricos) que fornece uma indicação do conteúdo da classe (WIPO, 2023).

Subclasse:

As classes são divididas em subclasses, as quais constituem o terceiro nível hierárquico da classificação. Subclasses são compostas pelo símbolo da subclasse, que consiste em uma letra maiúscula logo após o símbolo da classe, (por exemplo, H01S, onde “S” é o símbolo da subclasse) e pelo título da subclasse, que busca apresentar o conteúdo da subclasse com a maior precisão possível (WIPO, 2023). A subclasse H01S é definida por:

> H01S DISPOSITIVOS UTILIZANDO O PROCESSO DE AMPLIFICAÇÃO DA LUZ POR EMISSÃO ESTIMULADA DE RADIAÇÃO [LASER] PARA GERAR

> OU AMPLIFICAR LUZ; DISPOSITIVOS UTILIZANDO EMISSÃO ESTIMULADA DE RADIAÇÃO ELETROMAGNÉTICA EM FAIXAS DE FREQUÊNCIA OUTRAS QUE NÃO A ÓPTICA.

Grupo:

As subclasses são, ainda, divididas em grupos. Os grupos são organizados em grupos principais (quarto nível hierárquico da classificação) e subgrupos (níveis hierárquicos mais baixos que pertencem ao grupo principal). O símbolo do grupo é constituído da subclasse, seguido de dois números separados por um traço obliquo, por exemplo, H01S 3/00. O grupo que possui o número “00” logo após o traço obliquo é chamado de “grupo principal”. O título do grupo principal define de forma precisa um campo do assunto dentro de sua subclasse, por exemplo, H01S 3/00 Lasers (WIPO, 2023).

Os subgrupos são construídos da mesma forma que os grupos principais, se diferenciando apenas nos últimos números, os quais vêm após o traço obliquo, devendo ser diferentes de “00”, por exemplo, H01S 3/02. O título do subgrupo define um campo dentro do âmbito do seu grupo principal (WIPO, 2023).

Símbolo completo de classificação:

Um símbolo completo é composto de seção, classe, subclasse e grupo ou subgrupo (WIPO, 2023).

Exemplos:

1) A01B33/00
    a. A – Seção – 1º nível hierárquico
    b. 01 – Classe – 2º nível hierárquico
    c. B – Subclasse – 3º nível hierárquico
    d. 33/00 – Grupo principal – 4º nível hierárquico
2) A01B33/08
    a. A – Seção – 1º nível hierárquico
    b. 01 – Classe – 2º nível hierárquico
    c. B – Subclasse – 3º nível hierárquico
    d. 33/08 – Subgrupo – nível hierárquico mais baixo que grupo principal

Atualização da IPC e Reclassificação de Pedidos:

A OMPI realiza reuniões periódicas para melhorar e atualizar a estrutura da IPC. Sempre observando áreas do conhecimento em que a concentração de pedidos está

ou pode vir a aumentar. Anualmente, novos símbolos são criados para melhor especificar os documentos encontrados naquela área tecnológica, a fim de facilitar a busca de documentos.

Sempre que novos símbolos são criados, todos os pedidos que estavam classificados nos símbolos impactados precisam ser reclassificados. As alterações realizadas na IPC sempre precisam ser muito bem justificadas, pois cada símbolo impactado resulta em um número de pedidos que precisam ser reclassificados. Sendo a reclassificação um procedimento que normalmente é manual, gerando uma carga de trabalho para todos os países; quanto maior a quantidade de pedidos a serem reclassificados, maior a carga de trabalho.

Importância da Classificação:

Mesmo com muitos esforços para tentar realizar a classificação dos pedidos de forma automática, este é um procedimento que ainda tem uma atuação manual. Esse procedimento manual é possivelmente o único momento em que uma pessoa analisa o conteúdo técnico de um pedido de patente antes do exame propriamente dito.

A primeira ação de classificação pode ser o único evento, em anos, que faz com que o dito pedido fique melhor posicionado dentro da área do conhecimento a que ele pertence. Caso essa classificação seja de baixa qualidade, o pedido em questão ficará fora da área de conhecimento mais adequada ao seu conteúdo.

Muitos pedidos acabam nem sendo examinados pelos mais diversos motivos e, com isso, o conhecimento que se tornou público no momento da publicação irá ficar sempre fora do contexto correto no que diz respeito a classificação. Portanto, quanto melhor classificado, desde o início do processo, mais chances do pedido ser encontrado em uma eventual busca.

O pedido pode ser reclassificado a qualquer momento. Existem duas situações em que usualmente ocorrem as reclassificações: modificação no esquema de classificação; e durante o exame técnico.

Dados completos da IPC:

A OMPI disponibiliza, a cada versão, todos os arquivos necessários para a definição da IPC. Mesmo não sendo escopo deste trabalho, vale ressaltar que todos os textos da IPC podem ser fonte de informação para o treinamento de redes neurais. A IPC possui uma quantidade de informações muito grande e organizada podendo ser utilizada como fonte de dados para geração de algum tipo de modelo de rede neural. Este tipo de abordagem foi cogitado, mas no momento não será explorada[40] .

[40] Site da OMPI onde estão os arquivos que contém todas as informações do site da IPC. Disponível em: https://www.wipo.int/classifications/ipc/en/ITsupport/ Acesso em: 4 out, 2024.