UNIVERSIDADE NOVE DE JULHO
PROGRAMA DE PÓS-GRADUAÇÃO EM ADMINISTRAÇÃO - PPGA

KÁTIA CINARA TREGNAGO CUNHA

AQUISIÇÃO DE CONHECIMENTO NA *KNOWLEDGE-BASED VIEW*:
DESENVOLVIMENTO DE UM MÉTODO DE MINERAÇÃO DE PATENTES PARA APOIAR A CAPACIDADE ABSORTIVA ORGANIZACIONAL

São Paulo
2025

Kátia Cinara Tregnago Cunha

AQUISIÇÃO DE CONHECIMENTO NA *KNOWLEDGE-BASED VIEW*:

DESENVOLVIMENTO DE UM MÉTODO DE MINERAÇÃO DE PATENTES PARA APOIAR A CAPACIDADE ABSORTIVA ORGANIZACIONAL

KNOWLEDGE ACQUISITION IN A KNOWLEDGE-BASED VIEW:

DEVELOPMENT OF A PATENT MINING METHOD TO SUPPORT ORGANIZATIONAL ABSORPTIVE CAPACITY

TESE APRESENTADA AO PROGRAMA DE PÓS-GRADUAÇÃO EM ADMINISTRAÇÃO DA UNIVERSIDADE NOVE DE JULHO – UNINOVE, COMO REQUISITO PARA OBTENÇÃO DO GRAU DE **DOUTORA EM ADMINISTRAÇÃO**.

ORIENTADORA: PROFA. DRA. CRISTINA DAI PRÁ MARTENS
COORIENTADORA: PROFA. DRA. CARLA BONATO MARCOLIN
SUPERVISOR NO EXTERIOR: PROF. DR. CARLOS MANUEL JORGE DA COSTA (UNIVERSIDADE DE LISBOA, PORTUGAL)

São Paulo

2025

Cunha, Kátia Cinara Tregnago.
Aquisição de conhecimento na knowledge-based view: desenvolvimento de um método de mineração de patentes para apoiar a capacidade absortiva organizacional. / Kátia Cinara Tregnago Cunha. 2025.
269 f.
Tese (Doutorado)- Universidade Nove de Julho - UNINOVE, São Paulo, 2025.
Orientador (a): Profª. Drª. Cristina Dai Prá Martens.
1. Visão baseada no conhecimento. 2. Capacidade absortiva. 3. Patente. 4. Mineração textual. 5. Ontologia.
I. Martens, Cristina Dai Prá. II. Título

CDU 658

AQUISIÇÃO DE CONHECIMENTO NA *KNOWLEDGE-BASED VIEW*:
DESENVOLVIMENTO DE UM MÉTODO DE MINERAÇÃO DE PATENTES PARA APOIAR A CAPACIDADE ABSORTIVA ORGANIZACIONAL

POR

KÁTIA CINARA TREGNAGO CUNHA

Tese apresentada ao Programa de Pós-Graduação em Administração - PPGA da Universidade Nove de Julho – UNINOVE, como requisito para obtenção do título de Doutora em Administração, sendo a banca examinadora formada por:

Prof.(a) Dr (a). Cristina Dai Prá Martens (Orientadora)

Prof.(a) Dr (a). Carla Bonato Marcolin (Coorientadora- UFU)

Prof.(a) Dr (a) Cristiane Drebes Pedron (UNINOVE)

Prof.(a) Dr (a). Fernando Antonio Ribeiro Serra (UNINOVE)

Documento assinado digitalmente
gov.br RAQUEL JANISSEK MUNIZ
Data: 12/12/2025 16:34:27-0300
Verifique em https://validar.iti.gov.br

Prof.(a) Dr (a). Raquel Janissek-Muniz (UFRGS)

Documento assinado digitalmente
gov.br ADELAIDE MARIA DE SOUZA ANTUNES
Data: 12/12/2025 16:53:15-0300
Verifique em https://validar.iti.gov.br

Prof.(a) Dr (a). Adelaide Maria de Souza Antunes (INPI)

São Paulo, 12 de dezembro de 2025.

Aos meus pais, pelos valores que me transmitiram e por serem meus maiores exemplos e incentivadores na realização dos meus sonhos.
Ao meu esposo e ao meu filho, pelo amor e apoio que tornaram possível cada etapa desta caminhada.

Agradecimentos Institucionais

Agradeço à Uninove pela oportunidade de integrar uma instituição da qual me orgulho, pelo compromisso com a qualidade do ensino. Estendo meus agradecimentos a todos os professores que, com dedicação, contribuíram para minha formação.

À CAPES, pelo apoio fundamental ao desenvolvimento da pesquisa.

À Universidade de Lisboa, com especial gratidão ao Instituto Superior de Economia e Gestão (ISEG), por me receber durante o período de doutoramento sanduíche.

À UFRGS, minha casa profissional há mais de trinta anos, onde aprendi o verdadeiro valor da pesquisa e da educação pública.

Agradecimentos Pessoais

Agradeço às minhas orientadoras, Profa. Cristina e Profa. Carla, que foram essenciais nesta trajetória, por me desafiarem, inspirarem e, acima de tudo, por toda a parceria e apoio ao longo do caminho.

Ao Prof. Carlos J. Costa, sou grata pelos valiosos ensinamentos e pelas longas conversas acompanhadas de café, que tanto contribuíram para este trabalho e tornaram mais leve a estada em Portugal, longe da família.

Às minhas sempre orientadoras, Profa. Cristiane Drebes Pedron e Profa. Giandra Volpato, que desde o Mestrado me acompanham, incentivam e fazem parte desta conquista.

À Celise Marson, colega de pós-graduação e bibliotecária da Uninove, agradeço pelo apoio na busca e obtenção de artigos fundamentais para o desenvolvimento desta pesquisa.

Às queridas Ana Cândida, Daiane e Danielle, minhas colegas de doutorado e companheiras em Portugal, sou profundamente grata pela amizade, pelas risadas, pelos cafés, pelas taças de vinho, pelas idas ao Santuário de Fátima e pelo inesquecível Ano Novo em Coimbra.

Às chefias da UFRGS — em especial ao Prof. Helder Teixeira, à Profa. Karina Paese, ao Prof. Ruy Beck e ao Prof. Marcelo Arbo — agradeço pela confiança depositada em mim e pelo apoio contínuo ao longo deste percurso.

À colega Ana Jussara, pela parceria e apoio na condução das atividades na Faculdade de Farmácia/UFRGS.

Por fim, agradeço à minha família, minha base e meu porto seguro: ao meu pai, exemplo de dedicação e força; à minha mãe, que me acompanha de outro plano, pela presença constante e pela inspiração que ainda me oferece; e aos meus irmãos, por serem meu apoio incondicional. Ao meu esposo, agradeço por sua companhia em todos os momentos — pela paciência, compreensão e incentivo durante esta longa jornada. Ao meu filho, meu maior orgulho,

agradeço por me lembrar diariamente do sentido de tudo. À minha cunhada, pelo carinho e pela ajuda generosa ao longo do caminho.

Aos amigos, colegas e parceiros profissionais, deixo meu sincero agradecimento por compreenderem minhas ausências e por torcerem por mim, mesmo à distância. Estive, de fato, apaixonadamente dedicada a este trabalho, que hoje me enche de orgulho e que posso chamar de minha tese de doutorado.

“O conhecimento é uma obra coletiva”

Profa. Dra. Wrana Maria Panizzi

RESUMO

O conhecimento constitui um ativo estratégico fundamental das organizações, sendo determinante para a geração de vantagens competitivas sustentáveis, conforme postula a Visão Baseada no Conhecimento (*Knowledge-Based View* – KBV). Nessa perspectiva, a inovação emerge como resultado da capacidade organizacional de adquirir, assimilar, transformar e aplicar conhecimento, processo diretamente associado ao desenvolvimento da capacidade absortiva. A busca por informações científicas e tecnológicas, que compõem a chamada inteligência técnica, encontra nas bases de patentes um repositório de conhecimento codificado. Contudo, sua exploração é dificultada pelo elevado volume documental, pela heterogeneidade linguística e pela complexidade terminológica característica dos textos patentários. A mineração textual de patentes configura-se, assim, como uma tarefa desafiadora, porém essencial para viabilizar a aquisição sistemática dessa inteligência técnica, atuando como mecanismo habilitador da capacidade absortiva ao reduzir barreiras cognitivas, linguísticas e técnicas ao conhecimento externo. Nesse contexto, orientada teoricamente pela KBV e fundamentada metodologicamente no *Design Science Research* (DSR), esta pesquisa propõe-se a responder à seguinte questão central: como viabilizar a aquisição de inteligência técnica de patentes, de modo a permitir sua internalização e articulação com o conhecimento interno da organização, por meio de um método de mineração textual apoiado em uma ontologia? O objetivo geral consiste em propor um método de mineração textual, apoiado em ontologia semântica e linguística, capaz de viabilizar a aquisição, a assimilação e a transformação das informações técnicas contidas em patentes no contexto organizacional. A tese estrutura-se em quatro estudos interligados que consolidam o conhecimento sobre mineração de patentes, mapeiam sua evolução e identificam lacunas, com destaque para a integração entre TRIZ e mineração textual. A pesquisa é pioneira ao propor e validar um método híbrido de mineração textual de patentes que combina Inteligência Artificial, por meio de modelos de linguagem de grande escala, com uma lógica derivativa sustentada por uma ontologia semântica funcional baseada nos efeitos físicos da TRIZ, desenvolvida originalmente para a língua portuguesa. Os resultados demonstram que a articulação entre TRIZ, mineração textual e IA fortalece a capacidade absortiva organizacional e oferece uma base empírica e conceitual para transformar informação tecnológica em conhecimento estratégico, apoiando diretamente atividades de Pesquisa, Desenvolvimento e Inovação e ampliando o potencial competitivo de organizações e ecossistemas de inovação.

Palavras-chave: Visão Baseada no conhecimento. Capacidade absortiva. Patente. Mineração textual. Ontologia.

ABSTRACT

Knowledge constitutes a fundamental strategic asset of organizations, being decisive for the generation of sustainable competitive advantages, as postulated by the Knowledge-Based View (KBV). From this perspective, innovation emerges as the result of an organization's ability to acquire, assimilate, transform, and apply knowledge, a process directly associated with the development of absorptive capacity. The search for scientific and technological information, which composes what is known as technical intelligence, finds in patent databases a repository of codified knowledge. However, its exploitation is hindered by the large volume of documents, linguistic heterogeneity, and the terminological complexity characteristic of patent texts. Patent text mining thus constitutes a challenging yet essential task to enable the systematic acquisition of this technical intelligence, acting as an enabling mechanism of absorptive capacity by reducing cognitive, linguistic, and technical barriers to external knowledge. In this context, theoretically guided by the Knowledge-Based View and methodologically grounded in Design Science Research (DSR), this study seeks to answer the following central research question: how can the acquisition of technical intelligence from patents be enabled so as to allow its internalization and articulation with an organization's internal knowledge through a text-mining method supported by an ontology? The overall objective is to propose a text-mining method, supported by semantic and linguistic ontology, capable of enabling the acquisition, assimilation, and transformation of the technical information contained in patents within the organizational context. The thesis is structured into four interconnected studies that consolidate knowledge on patent mining, map its evolution, and identify gaps, with particular emphasis on the integration between TRIZ and text mining. The research is pioneering in proposing and validating a hybrid patent text-mining method that combines Artificial Intelligence, through large language models, with a derivative logic supported by a functional semantic ontology based on TRIZ physical effects, originally developed for the Portuguese language. The results demonstrate that the articulation between TRIZ, text mining, and AI strengthens organizational absorptive capacity and provides an empirical and conceptual basis for transforming technological information into strategic knowledge, directly supporting Research, Development, and Innovation activities and enhancing the competitive potential of organizations and innovation ecosystems.

Keywords: *Knowledge-Based View. Absorptive Capacity. Patent. Text Mining. Ontology.*

LISTA DE ILUSTRAÇÕES

FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ARIPO	*African Regional Intellectual Property Organization*
BERT	*Bidirectional Encoder Representations from Transformers*
BLEU	*Bilingual Evaluation Understudy*
BoW	*Bag-of-Words*
CNN	Rede Neural Convolucional
CorEx	*Correlation Explanation*
CPC	Classificação Cooperativa de Patentes
CPLP	Comunidade dos Países de Língua Portuguesa
DL	*Deep Learning*
DSR	*Design Science Research*
EAPO	*Eurasian Patent Organization*
EF	Efeito físico
EPO	*European Patent Office*
EM	*Exact match*
FI	Fator de impacto
FP	Falso Positivo
GAN	*Generative Adversarial Network*
GTM	*Generative Topographic Map*
IA	Inteligência Artificial
IBICT	Instituto Brasileiro de Informação em Ciência e Tecnologia
IDM	*Inventive Design Method*
INPI	Instituto Nacional da Propriedade Industrial
IPC	*International Patent Classification*
JCR	*Journal Citation Report*
KBV	*Knowledge-Based View*
LDA	*Latent Dirichlet Allocation*
LLM	*Large Language Model*
LSA	*Latent Semantic Analysis*
LSTM	*Long Short-Term Memory*
ML	*Machine Learning*
MT	Mineração Textual

NEN	*Named Entity Normalization*
NER	*Named Entity Recognition*
NLP	*Natural Language Processing*
NLTK	*Natural Language Toolkit*
NPMI	*Normalized Pointwise Mutual Information*
O	Objeto
OAPI	*Organisation Africaine de la Propriété Intellectuelle*
ODS	Objetivos de Desenvolvimento Sustentável
ONU	Organização das Nações Unidas
ORSD	*Ontology Requirements Specification Document*
OWL	*Ontology Web Language*
P&D	Pesquisa e Desenvolvimento
PD&I	Pesquisa, Desenvolvimento e Inovação
PF	Propriedade-função
PLN	Processamento de Linguagem Natural
PLsA	*Probabilistic Latent Semantic Analysis*
PLSi	*Probabilistic Latent Semantic Indexing*
POS	*Part-of-Speech*
PRISMA	*Preferred Reporting Items for Systematic Reviews and Meta-Analyses*
PTT	Produto Técnico-Tecnológico
RI	Recuperação de Informação
ROUGE	*Recall-Oriented Understudy for Gisting Evaluation*
RSL	Revisão Sistemática da Literatura
SAO	Sujeito-Ação-Objeto
StArt	*State of the Art through Systematic Review*
SVD	*Singular Value Decomposition*
T	Tarefa
TF-IDF	*Term Frequency–Inverse Document Frequency*
TRIZ	Teoria da Resolução Inventiva de Problemas
URL	*Uniform Resource Locator*
USPTO	Escritório de Marcas e Patentes dos Estados Unidos
VP	Verdadeiro Positivo
WIPO	*World Intellectual Property Organization*

SUMÁRIO

1 INTRODUÇÃO

Em ambientes caracterizados por rápidas mudanças tecnológicas e competitivas, o conhecimento tem se consolidado como um dos principais determinantes do desempenho organizacional e da inovação. Mais do que um insumo operacional, o conhecimento orienta a capacidade das empresas de aprender, adaptar-se e renovar continuamente suas bases de recursos, sustentando vantagens competitivas ao longo do tempo (Grant, 1996; Zheng et al., 2011).

Nesse contexto, as organizações buscam continuamente adquirir, gerar e recombinar conhecimentos internos e externos como forma de responder às transformações do ambiente. Tais processos envolvem mecanismos de criação, compartilhamento, integração e gestão do conhecimento (Stoian et al., 2024), que permitem às empresas reconfigurar suas competências e desenvolver novas capacidades.

É a partir dessa centralidade do conhecimento que a Visão Baseada no Conhecimento (*Knowledge-Based View* – KBV) emerge, ao estabelecer o conhecimento como o recurso estratégico fundamental das organizações (Takeuchi, 2013). Sob essa perspectiva, a habilidade de mobilizar, integrar e transformar diferentes tipos de conhecimento torna-se relevante para a construção e sustentação de vantagens competitivas (Cabrilo & Dahms, 2018; Grant, 1996, 1997; Grant & Phene, 2022), sendo operacionalizada por meio de capacidades dinâmicas baseadas em processos de aprendizagem e gestão do conhecimento.

Para alimentar seu estoque de recursos, as empresas recorrem à aquisição de conhecimento externo por meio de processos de aprendizagem organizacional. Uma vez acessado, esse conhecimento é processado pela capacidade absortiva, entendida como a habilidade da empresa de identificar, assimilar, transformar e combinar novos conhecimentos com aqueles já existentes (Zahra & George, 2002), possibilitando a geração de novos produtos, processos e capacidades organizacionais (Cohen & Levinthal, 1990). Esse conhecimento pode ser obtido por meio de investimentos em pesquisa e desenvolvimento (P&D), processos de contratação, criação de rotinas de aprendizagem, estímulo ao compartilhamento do conhecimento e adoção de estratégias de inovação aberta.

A KBV e a Teoria da Capacidade Absortiva concentram-se predominantemente no processamento do conhecimento após seu ingresso na organização. Contudo, fatores externos podem impedir o reconhecimento ou o acesso ao conhecimento externo, tornando-o invisível ou inalcançável para a firma antes mesmo que qualquer esforço de assimilação possa ocorrer.

As barreiras antecedentes à aquisição do conhecimento podem ser de natureza cognitiva,

relacional ou estrutural. As barreiras cognitivas referem-se à incapacidade da empresa de "enxergar" ou valorizar o conhecimento externo (Nooteboom, 2000; Katz & Alen, 1982; Levinthal & March, 1993). Nooteboom (2000) argumenta que, quando o conhecimento externo é excessivamente distante da base cognitiva da organização, esta carece dos esquemas mentais necessários para reconhecê-lo. Katz e Allen (1982) descrevem a Síndrome do Não Inventado Aqui (*Not-Invented-Here*) onde grupos técnicos, com o passar do tempo, tendem a rejeitar conhecimento externo decorrente da crença de que as fontes externas não têm nada de novo a ensinar, sendo visto como uma ameaça ao *status* e à competência do grupo. Levinthal e March (1993), por sua vez, introduzem o conceito de miopia organizacional, destacando como as empresas tendem a aprender com o que está próximo (vizinhos, parceiros atuais, tecnologias familiares) e ignoram o que está longe ou em domínios diferentes, explorando as competências atuais e, assim, ignorando novas fontes de conhecimento.

As barreiras de natureza relacional dizem respeito à posição da empresa em redes sociais e à dificuldade de acesso ao conhecimento em função da ausência ou fragilidade de conexões. Granovetter (1973) argumenta que redes excessivamente fechadas, baseadas em laços fortes (parcerias de longa data), podem limitar a entrada de informações novas porque não há pontes para o mundo exterior. Uzzi (1997) discute como a imersão em redes sociais pode tanto facilitar quanto isolar as empresas de fontes externas de conhecimento. Hansen (1999) destaca, ainda, as dificuldades relacionais associadas à identificação de fontes de conhecimento úteis em organizações grandes e complexas. A ausência de pontes (laços fracos) que conectem a unidade aos diversos depósitos de conhecimento da organização complexa torna as fontes de conhecimento invisíveis para quem precisa delas, levando a equipe a utilizar apenas o conhecimento interno (conectando-se à ideia de Miopia de Levinthal e March (1993)).

Por fim, as barreiras de natureza estrutural e legal referem-se a impedimentos formais, geográficos ou de custo que bloqueiam o acesso ao conhecimento antes mesmo de sua aquisição. Teece (1986) introduz o conceito de regime de apropriabilidade, explicando que patentes e mecanismos de proteção legal funcionam como barreiras estruturais para aqueles que buscam acessar determinado conhecimento. Laursen e Salter (2006) acrescentam que a ausência de canais estruturados de inovação aberta também pode impedir que o conhecimento externo alcance a firma.

Essas barreiras tornam-se particularmente relevantes quando o conhecimento externo está formalizado em sistemas complexos e codificados, cujo acesso exige não apenas disponibilidade institucional, mas também competências cognitivas, rotinas organizacionais e conexões adequadas para sua exploração efetiva. Nesse contexto, as patentes configuram-se

como uma importante fonte de conhecimento técnico formalmente acessível por meio de bases públicas, como *Espacenet*, *Patentscope*, *The Lens*, *Google Patents* e os bancos de dados de autoridades nacionais, representados, no Brasil, pelo Instituto Nacional da Propriedade Industrial (INPI). Todavia, apesar dessa ampla disponibilidade, diversos estudos ao longo do tempo indicam o uso limitado dessas bases como recurso de conhecimento por parte da sociedade, de empresas, de formuladores de políticas públicas e de pesquisadores. A Figura 1 apresenta uma cronologia da subutilização de patentes documentada na literatura acadêmica.

Figura 1

Linha do tempo das pesquisas sobre limitações no uso de informações de patentes

Nota. Cunha et al. (2023); Marmor et al. (1979); Masurel (2005); Mazieri et al. (2016); McTeague e Chatzimichali (2022); Pimenta (2017); Rogers et al. (2012).

Os motivos para o uso limitado das informações contidas em patentes são variados. Os documentos patentários são frequentemente percebidos como longos e complexos (Donald et al., 2018; Suominen et al., 2018; Zhang et al., 2018), ricos em terminologia técnica e jurídica (Tseng et al., 2007; Xie & Miyazaki, 2013) e com estrutura e forma de expressão distintas das utilizadas em artigos científicos (Cunha et al., 2023). Para a prospecção nas bases patentárias, consideradas "enciclopédias técnicas de uso não óbvio" (Reymond & Quoniam, 2016, p. 5), são requeridas habilidades muitas vezes não contempladas na formação profissional (Quintella et al., 2011). Além disso, a quantidade de documentos disponíveis nessas bases supera amplamente a capacidade humana de leitura (Chiarello et al., 2018).

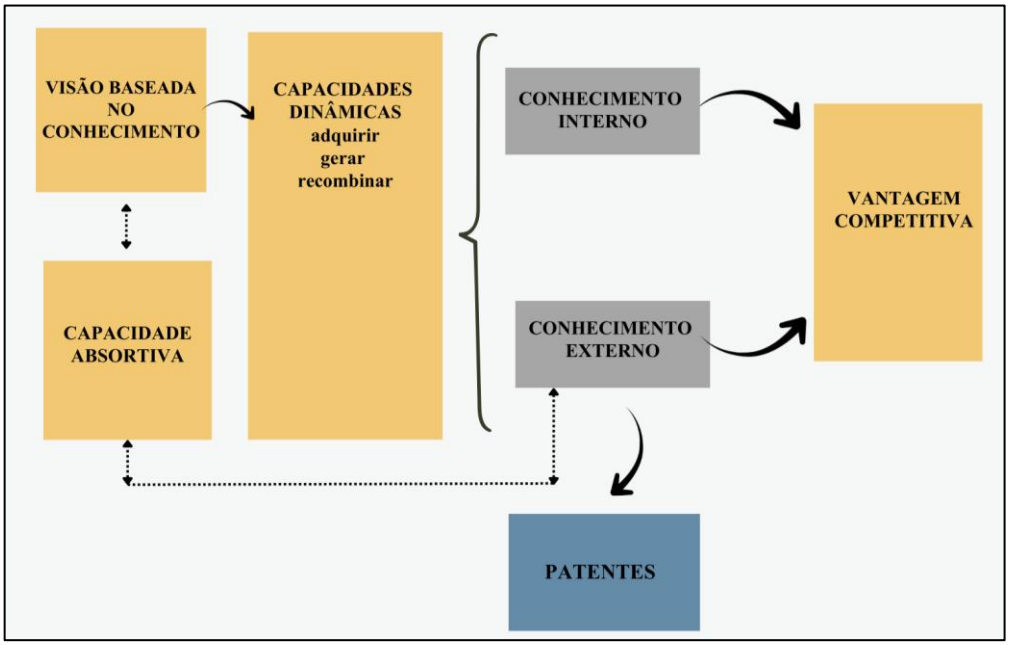
Essas questões configuram barreiras estruturais e cognitivas que impactam o uso do conhecimento externo oriundo de patentes. Embora as patentes sejam tradicionalmente reconhecidas como instrumentos jurídicos de proteção de ativos intangíveis, elas também constituem ativos estratégicos de aprendizagem, capazes de fornecer inteligência técnica recente e avançada em uma ampla variedade de domínios tecnológicos, inclusive em contextos que envolvem interações entre diferentes subdomínios do conhecimento (Deng & Wang, et al., 2018; Krestel & Chikkamath, et al., 2021).

A inteligência técnica compreende informações de natureza técnica e tecnológica (Behkami & Daim, 2012) as quais estruturam atributos tecnológicos por meio da linguagem para descrever tecnologias voltadas à resolução de problemas técnicos (Yoon & Park, 2004). A informação técnica, associada a recursos intangíveis, revela conhecimentos, habilidades e ideias mobilizados pelos gestores nas atividades organizacionais (Ali et al., 2020). Em contraste, a informação tecnológica está relacionada a recursos tangíveis, manifestando-se na forma de produtos e processos (Jang et al., 2021).

A inteligência técnica contida nas patentes, quando combinada com experiência, contexto, interpretação e reflexão (Jarrar, 2002), transforma-se em conhecimento. Esse conhecimento impacta diretamente as capacidades dinâmicas das organizações, fornecendo subsídios para a tomada de decisão, o fortalecimento da competitividade (Bianchi et al., 2016) e o fomento aos processos de inovação (He et al., 2022; Robert & Mayer, 2003). A Figura 2 apresenta a relação dinâmica entre a Visão Baseada no Conhecimento (KBV), a Capacidade Absortiva e as Patentes.

Figura 2

Patentes como fonte de informação técnica: implicações para as capacidades dinâmicas na Visão Baseada no Conhecimento (KBV) e Capacidade Absortiva



Nota. Elaborado pela Autora (2025).

De forma integrada, o modelo evidencia que a KBV fundamenta a compreensão de que o conhecimento constitui o principal recurso estratégico da organização. Por meio da Capacidade Absortiva, o conhecimento externo proveniente de documentos de patente é internalizado e articulado ao conhecimento interno já existente, sendo transformado e explorado pelas capacidades dinâmicas. As patentes, enquanto fonte estruturada de conhecimento externo, fornecem informações técnicas relevantes que alimentam um processo contínuo e cumulativo de aprendizagem organizacional.

1.1 Problema de Pesquisa

As patentes constituem uma fonte pública, estruturada e estrategicamente relevante de conhecimento técnico externo, amplamente reconhecida por seu potencial de apoiar processos de aprendizagem organizacional, inovação e desenvolvimento tecnológico (Liwei, 2022; Pimenta, 2017; Xu et al., 2022). Todavia, apesar de sua disponibilidade institucional e de seu valor informacional, observa-se um paradoxo persistente: as patentes permanecem subutilizadas como fonte efetiva de inteligência técnica pelas organizações.

Esse paradoxo decorre da existência de barreiras antecedentes à aquisição do conhecimento, que se manifestam antes mesmo dos processos tradicionalmente analisados pela

literatura de Capacidade Absortiva. Tais barreiras são de natureza cognitiva, estrutural, metodológica e linguística, limitando o reconhecimento, o acesso e a interpretação do conhecimento contido nos documentos patentários e, consequentemente, inviabilizando sua internalização e articulação ao conhecimento organizacional (Cunha et al., 2023; Masurel, 2005; Mazieri et al., 2016; McTeague & Chatzimichali, 2022; Pimenta, 2017; Rogers et al., 2012).

Do ponto de vista cognitivo, os documentos de patente diferem substancialmente de outros tipos de textos técnicos ou científicos, tanto em sua estrutura quanto em sua lógica discursiva (Chen et al., 2022). Essa distinção torna a recuperação de informações relevantes uma tarefa particularmente desafiadora (Florescu & Caragea, 2017; Liwei, 2022), exigindo elevado esforço interpretativo e intensa intervenção manual (Hu et al., 2018). A complexidade é amplificada pela coexistência, em um mesmo documento, de linguagem técnica, jurídica e estratégica, além do uso deliberado de variações semânticas e formulações não triviais, inerentes à redação patentária (Fall et al., 2003; Liwei, 2022).

Além disso, os textos de patentes são estruturados em torno da descrição de problemas técnicos e das soluções propostas para resolvê-los, refletindo uma lógica inventiva específica que nem sempre é capturada por métodos linguísticos tradicionais. Essa lógica foi sistematizada pela Teoria da Resolução Inventiva de Problemas (TRIZ), desenvolvida a partir da análise em larga escala de patentes e fundamentada na identificação de padrões recorrentes de resolução de contradições técnicas. A ausência de estruturas conceituais alinhadas a essa lógica inventiva contribui para a dificuldade das organizações em reconhecer, interpretar e explorar o conhecimento técnico contido nos documentos patentários.

Do ponto de vista metodológico, técnicas tradicionais de mineração textual e extração de termos mostram-se limitadas para lidar com essa variedade semântica, estrutural e inventiva. Essas abordagens frequentemente não conseguem identificar os conceitos mais representativos das invenções descritas, especialmente aqueles relacionados a problemas técnicos, contradições e princípios de solução, comprometendo a precisão da extração, da classificação e do processamento do conhecimento técnico (Kim et al., 2008; Berdyugina & Cavallucci, 2023; Kaliteevskii et al., 2021; Kang et al., 2018; Kim et al., 2019). Como consequência, o conhecimento potencialmente disponível nas patentes permanece fragmentado ou inacessível para os tomadores de decisão organizacionais.

Esse cenário evidencia uma lacuna teórica e metodológica relevante na literatura: embora a Visão Baseada no Conhecimento e a Teoria da Capacidade Absortiva expliquem como o conhecimento externo é assimilado, transformado e explorado após sua aquisição, elas

oferecem explicações limitadas sobre como esse conhecimento se torna visível**,** acessível e cognitivamente processável antes de sua assimilação**,** especialmente quando formalizado em sistemas complexos como as bases patentárias.

Diante disso, emerge a necessidade de métodos capazes de reduzir as barreiras antecedentes à aquisição do conhecimento, estruturando, organizando e traduzindo a inteligência técnica contida nas patentes de forma compatível com os esquemas cognitivos organizacionais. Nesse contexto, abordagens baseadas em ontologias, especialmente quando informadas por estruturas conceituais da TRIZ, apresentam-se como promissoras, por permitirem a representação formal do conhecimento, a desambiguação semântica e a organização conceitual de domínios tecnológicos orientados à resolução de problemas, potencializando processos automatizados de mineração textual e recuperação de informação.

Dessa forma, o problema de pesquisa que orienta esta tese pode ser formulado da seguinte maneira: Como viabilizar a aquisição de inteligência técnica de patentes, de modo a permitir sua internalização e articulação com o conhecimento interno da organização, por meio de um método de mineração textual apoiado em uma ontologia?

1.2 Objetivos

A presente tese está definida em um objetivo geral e objetos específicos, os quais serão a seguir apresentados.

1.2.1 Objetivo Geral

Propor um método de mineração textual, apoiado em ontologia semântica e linguística, para viabilizar a aquisição e internalização das informações técnicas contidas em patentes no contexto organizacional.

1.2.2 Objetivos Específicos

I. Mapear metodologias de mineração de textos aplicadas a documentos de patentes.
II. Analisar a tendência metodológica dos métodos de mineração textual que integram os conceitos e ferramentas da Teoria da Resolução Inventiva de Problemas (TRIZ).
III. Construir uma ontologia formal, fundamentada nos efeitos físicos da TRIZ, que funcione como base semântica e linguística para a extração automática de soluções genéricas a partir de documentos de patentes.

IV. Desenvolver e validar um método de mineração de inteligência técnica em documentos de patentes, utilizando a ontologia como base semântica para aplicações em linguística computacional e processos de inteligência artificial.

1.3 Justificativa

A informação técnica constitui um insumo crítico para a geração, combinação e reconfiguração de bases de recursos organizacionais, sendo igualmente fundamental para a constituição da infraestrutura necessária à inovação (Robert & Mayer, 2003; S. Zheng et al., 2011). Nesse contexto, as bases patentárias destacam-se como fontes estratégicas de informação tecnológica, pois reúnem um volume expressivo, continuamente atualizado e, em muitos casos, exclusivo de soluções técnicas já desenvolvidas (Aristodemou et al., 2017; de Weck, 2022; Siddharth et al., 2022).

Apesar de sua relevância, estudos históricos e contemporâneos indicam que as bases de patentes permanecem subutilizadas por pesquisadores e organizações (Bregonje, 2005; Marmor et al., 1979; Pimenta, 2017). Evidências recentes obtidas a partir de entrevistas exploratórias com pesquisadores brasileiros revelam que menos da metade utiliza essas bases de forma sistemática, sendo apontadas como principais barreiras a complexidade dos textos, a linguagem técnico-jurídica e a baixa familiaridade com o sistema de propriedade intelectual (Cunha et al., 2023). Essas dificuldades são agravadas pela sobrecarga informacional e pela necessidade de recursos humanos altamente especializados para interpretar os documentos, especialmente diante da escala das bases disponíveis — como a Espacenet, a Patentscope e o banco de dados do INPI, que, conjuntamente, concentram centenas de milhões de documentos (European Patent Office, 2025a; INPI, 2024; World Intellectual Property Organization, 2025b).

Sob a perspectiva da KBV, os documentos de patentes configuram-se como ativos de conhecimento codificado. Entretanto, seu valor estratégico não reside apenas no acesso à informação neles contida, mas sobretudo na capacidade organizacional de transformá-la em conhecimento acionável, passível de integração aos processos de P&D, inovação e tomada de decisão estratégica. Nesse contexto, os métodos de mineração textual atuam como mecanismos intermediários essenciais, ao viabilizar a conversão de informação técnica dispersa em conhecimento estruturado, reduzindo barreiras cognitivas e técnicas à sua exploração.

A ampliação do acesso e do uso estratégico das informações técnicas contidas em patentes contribui para a promoção de avanços tecnológicos, fortalecimento da pesquisa

científica e o impulsio à inovação (Audretsch & Feldman, 1996; Belenzon, 2012; Han et al., 2006). No contexto empresarial, especialmente em pequenas e médias empresas, esse conhecimento pode subsidiar a identificação de oportunidades de mercado, orientar estratégias tecnológicas e aumentar a competitividade em ambientes caracterizados por rápida evolução tecnológica e intensa concorrência global (He et al., 2022; Liu et al., 2016). A Figura 3 apresenta as principais contribuições da informação técnica obtida de patentes para o fortalecimento da capacidade inovadora e competitiva das organizações.

Figura 3

Inteligência técnica de patentes como vantagem competitiva e de inovação

Nota. Elaborado pela Autora (2025).

Nesse cenário, a presente pesquisa dialoga diretamente com os princípios da Agenda 2030 das Nações Unidas, ao reconhecer o papel das patentes como instrumentos de disseminação de tecnologias inovadoras capazes de contribuir para o desenvolvimento sustentável. Destaca-se, em especial, o alinhamento com o Objetivo de Desenvolvimento Sustentável (ODS) 9 (Indústria, Inovação e Infraestrutura) por meio da proposição de um método de mineração textual voltado à extração de informação técnica de patentes, fortalecendo

infraestruturas de pesquisa e ampliando o acesso ao conhecimento tecnológico. De forma complementar, a pesquisa contribui para o ODS 4 (Educação de Qualidade), ao oferecer metodologias aplicáveis à formação de profissionais em áreas como engenharia, ciência da informação e inovação; para o ODS 8 (Trabalho Decente e Crescimento Econômico), ao apoiar a geração de novos produtos e serviços baseados em conhecimento; e para os ODS 16 e 17, ao fomentar transparência, cooperação e parcerias entre universidades, empresas e governos. A Figura 4 sumariza a contribuição da pesquisa para o alcance dos ODS.

Figura 4

Contribuição da pesquisa para o alcance dos Objetivos de Desenvolvimento Sustentável

Nota. Elaborado pela Autora (2025).

Assim, ao propor soluções metodológicas para a mineração de informação técnica em documentos de patentes, esta investigação não apenas avança o estado da arte no campo técnico-científico, mas também reforça a relevância social, econômica e estratégica do conhecimento produzido, contribuindo para a construção de ecossistemas de inovação mais inclusivos, sustentáveis e competitivos.

1.4 Knowledge-Based View e Capacidade Absortiva Organizacional

A Visão Baseada no Conhecimento (KBV) emerge como um desdobramento da Visão Baseada em Recursos (*Resource-Based View* – RBV), ao deslocar o foco analítico dos recursos tangíveis para o conhecimento como o principal ativo estratégico das organizações. Sob a KBV, o conhecimento é compreendido como um recurso heterogêneo, difícil de imitar, socialmente complexo e cumulativo, cuja criação, integração e aplicação sustentam a vantagem competitiva ao longo do tempo (Grant, 1996; Kogut & Zander, 1996; Nonaka, 1994).

Grant (1996)) argumenta que a firma existe, fundamentalmente, como um mecanismo de integração de conhecimentos especializados dispersos entre indivíduos. Nessa perspectiva, o desempenho organizacional não depende apenas da posse de conhecimento, mas, sobretudo, da capacidade da organização de coordenar, combinar e aplicar diferentes tipos de conhecimento (tácito e explícito) de maneira eficiente. A KBV, portanto, enfatiza processos organizacionais de aprendizagem, comunicação, coordenação e rotinização como elementos centrais da estratégia.

É nesse contexto que a Capacidade Absortiva se consolida como um dos principais mecanismos explicativos da KBV. Introduzida por Cohen e Levinthal (1990), a capacidade absortiva é definida como a habilidade da organização de reconhecer o valor de novos conhecimentos externos, assimilá-los e aplicá-los para fins comerciais. Os autores destacam que essa capacidade é fortemente dependente do estoque prévio de conhecimento da firma, conferindo-lhe um caráter cumulativo e moldado pelo que a empresa aprendeu no passado.

Zahra e George (2002) ampliam e refinam essa conceituação ao definir a capacidade absortiva como um conjunto de rotinas e processos organizacionais por meio dos quais as empresas adquirem, assimilam, transformam e exploram o conhecimento. Essa reconceitualização introduz a distinção entre capacidade absortiva potencial, composta pelos processos de aquisição e assimilação, e capacidade absortiva realizada, formada pelos processos de transformação e exploração. Tal distinção tornou-se central na literatura ao evidenciar que a simples aquisição de conhecimento externo não garante, por si só, a geração de valor organizacional.

Sob a ótica da KBV, a capacidade absortiva pode ser entendida como uma capacidade dinâmica essencial, pois permite à organização reconfigurar continuamente sua base de conhecimento em resposta a mudanças ambientais (Grant & Phene, 2022; Teece,Pisano & Shuen et al., 1997). Ao integrar conhecimento externo com o conhecimento interno existente,

a organização amplia sua base cognitiva, fortalece suas rotinas e desenvolve novas competências, sustentando processos de inovação incremental e, em alguns casos, radical.

A literatura empírica demonstra consistentemente que a capacidade absortiva está positivamente associada à inovação, ao desempenho organizacional e à vantagem competitiva (Cabrilo & Dahms, 2018; Lane et al., 2006; Pu & Liu, 2023). Estudos recentes reforçam que a complementaridade entre capacidade absortiva potencial e realizada é fundamental, sendo que organizações que conseguem equilibrar esses dois componentes apresentam melhores resultados em termos de inovação e desempenho (Stettler et al., 2024; Zahra & George, 2002).

Adicionalmente, pesquisas contemporâneas têm enfatizado a natureza multinível da capacidade absortiva, reconhecendo que ela emerge da interação entre capacidades individuais, práticas organizacionais e estruturas institucionais (Xiong et al., 2024). Esse entendimento reforça a centralidade da KBV ao demonstrar que o conhecimento não reside apenas em artefatos ou documentos, mas é socialmente construído e operacionalizado por meio de indivíduos, equipes e rotinas organizacionais.

Nesse sentido, a capacidade absortiva funciona como um elo fundamental entre a disponibilidade de conhecimento externo e sua efetiva conversão em valor estratégico. Enquanto a KBV estabelece o conhecimento como o recurso estratégico central da firma, a capacidade absortiva explica como esse conhecimento é identificado, internalizado, transformado e explorado. Assim, ambos os constructos se complementam teoricamente, oferecendo um arcabouço robusto para a compreensão dos processos de aprendizagem organizacional, inovação e desenvolvimento de capacidades dinâmicas em ambientes intensivos em conhecimento

1.5 Estrutura da Tese

A tese está estruturada em quatro estudos interdependentes e interligados, derivados de um mesmo problema de pesquisa, conforme proposto por Costa et al. (2019). Cada estudo corresponde a, pelo menos, uma etapa do método de pesquisa adotado, o *Design Science Research* (DSR), com orientação teórica fundamentada na KBV.

No **Estudo 1** foi realizada uma Revisão Sistemática da Literatura (RSL) com o objetivo de investigar de que forma os métodos de mineração de texto realizam a extração e a análise de dados não estruturados provenientes de documentos de patente. Esse estudo identificou as principais adversidades enfrentadas pelos sistemas de recuperação de informação, decorrentes

das peculiaridades estruturais e semânticas dos textos e do grande volume de dados envolvidos (Chen et al., 2020; Fall et al., 2003). Também se observou a predominância de métodos de mineração voltados a linguagens de alto recurso e uma tendência de pesquisa direcionada à exploração de abordagens semânticas dos textos de patente, visando extrair soluções para problemas técnicos específicos com o apoio das ferramentas da TRIZ.

Com o objetivo de aprofundar a compreensão sobre a aplicação das ferramentas da TRIZ em tarefas de mineração textual de patentes, o **Estudo 2** consistiu em uma nova RSL, voltada à identificação de métodos de recuperação de inteligência técnica em documentos de patente baseados na TRIZ. Os resultados demonstram que as ferramentas da TRIZ oferecem um conjunto de conceitos fundamentais, relacionados a parâmetros, propriedades e funções, bastante semelhante aos termos empregados nos processos inventivos descritos em patentes. Esse conjunto, quando organizado em categorias, forma um *corpus* anotado inicial para tarefas de recuperação de informações, o que potencializa o uso de ferramentas computacionais.

Diante da inexistência de *corpora* anotados em língua portuguesa voltados à mineração de textos de patente, no **Estudo 3** foi desenvolvida uma ontologia semântica baseada nos efeitos físicos da TRIZ para apoiar a mineração de textos de patentes redigidos em português. A ontologia apresenta 11.196 entradas, estruturadas como um relacionamento ternário que integra as subclasses Tarefa (T), Objeto (O) e Efeito Físico (EF), termos semanticamente relacionados ao vocabulário técnico-patentário.

Por fim, no **Estudo 4**, foi desenvolvido um método de mineração textual de patentes, utilizando a ontologia como base linguística para a extração de inteligência técnica. Durante o desenvolvimento, foram testadas técnicas clássicas de Processamento de Linguagem Natural (PLN) e arquiteturas baseadas em Modelos de Linguagem de Grande Escala (LLMs), culminando na escolha de uma abordagem híbrida: LLMs com um módulo analítico auxiliar que, mediante regras, decide adotar os componentes semânticos Tarefa, Objeto e Efeito Físico extraídos do campo textual analisado ou utilizar os componentes semânticos da ontologia. A validação dos termos extraídos foi conduzida por sete especialistas, alcançando uma performance global de 73,26% (Precisão e *Exact Match*), o que posiciona o método de forma competitiva em relação a abordagens recentes (Blume et al., 2024; Lee & Bai, 2025; Li, Yu, et al., 2023; Miric et al., 2023; Trapp & Warschat, 2025).

Para visualizar a integração e a interconexão dos trabalhos, a Tabela 1 apresenta a Matriz Metodológica de Amarração (MMA), um diagrama que sintetiza de forma integrada os quatro estudos desenvolvidos ao longo desta tese.

Tabela 1

Matriz Metodológica de Amarração

QUESTÃO PRINCIPAL DE PESQUISA: Como viabilizar a aquisição de inteligência técnica de patentes, de modo a permitir sua internalização e articulação com o conhecimento interno da organização, por meio de um método de mineração textual apoiado em uma ontologia?					
OBJETIVO PRINCIPAL: Propor um método de mineração textual, apoiado em ontologia semântica e linguística, para viabilizar a aquisição e internalização das informações técnicas contidas em patentes no contexto organizacional.					
PARADIGMA DA TESE: Pragmatista					
JUSTIFICATIVA DE DISTINÇÃO		**JUSTIFICATIVA DE INTERDEPENDÊNCIA ESTRUTURAL E DE CONVERGÊNCIA CIENTÍFICA, TÉCNICA, TECNOLÓGICA E/OU SOCIAL**			
Campo de pesquisa ou título	**Objetivos específicos**	**Método único ou misto nas etapas da pesquisa**	**Procedimento de análise e coleta de dados**	**Produtos científicos e impactos potenciais**	**Produtos técnicos, tecnológicos e/ou sociais e impactos potenciais**
REVISÃO SISTEMÁTICA DA LITERATURA SOBRE MINERAÇÃO DE CAMPOS TEXTUAIS DE DOCUMENTOS DE PATENTE	Mapear metodologias de mineração de textos aplicadas a documentos de patentes.	Método único: Revisão sistemática da Literatura nas bases *Scopus* e *Web of Science*	•Estrutura PRISMA para condução da RSL •Plataforma Rayyan para a seleção dos artigos •Análise estatística das informações bibliográficas das publicações	Artigo científico publicado no Management Review Quarterly (ABS 1, CAPES MB, JCR Q1)	**Impacto acadêmico** - consolida pesquisas, sintetiza estratégias, identifica lacunas e orienta futuras agendas de pesquisa na mineração textual de patentes. **Impacto prático** - consolida conhecimento validado para pesquisas futuras.
INTEGRAÇÃO ENTRE TRIZ E MINERAÇÃO DE TEXTOS DE PATENTES: AVANÇOS, DESAFIOS E TENDÊNCIAS NA EXTRAÇÃO DE INTELIGÊNCIA TÉCNICA	Analisar a tendência metodológica dos métodos de mineração textual que integram os conceitos e ferramentas da TRIZ.	Método único: Revisão Sistemática da Literatura na *Base Scopus*	•Estrutura PRISMA para condução da RSL •Ferramenta *VOSviewer* para a análise de coocorência e inter-relação de termos •Análise estatística básica das informações	Artigo científico submetido ao *World Patent Information* (CAPES B, JCR Q2)	**Impacto acadêmico** - oferece uma visão abrangente sobre metodologias existentes, avanços e desafios na integração das técnicas de mineração de texto com as ferramentas TRIZ. **Impacto prático** –sintetiza os estudos e ferramentas de processamento linguístico específicos para a mineração textual de patente em associação com os conceitos da TRIZ, oferecendo um roteiro para

			bibliográficas das publicações		subsidiar o desenvolvimento de aplicações práticas.
DESENVOLVIMENTO DE UMA ONTOLOGIA BASEADA NA TEORIA DA SOLUÇÃO INVENTIVA DE PROBLEMAS (TRIZ)	Construir uma ontologia formal, fundamentada nos efeitos físicos da TRIZ, que funcione como base semântica e linguística para a extração automática de soluções genéricas a partir de documentos de patente.	Metodologia NeOn, com os requisitos definidos por meio de um Documento de Especificação de Requisitos de Ontologia Ferramenta de código aberto Protégé/Stanford para a construção da rede semântica	Termos obtidos do Banco de Dados Multilíngue de Efeitos Físicos, proposto por Zaniro et al. (2024)	Artigo completo publicado nos Anais do *13th World Conference on Information Systems and Technologie (Lecture Notes in Networks and Systems, Springer),* em colaboração com pesquisador do Instituto Superior de Economia e Gestão da Universidade de Lisboa.	**Produto técnico-tecnológico**[1]: • Base de dados técnico-científica (ontologia) **Impacto acadêmico** – Desenvolvimento de uma ontologia semântica funcional baseada em efeitos físicos da TRIZ, atuando como uma base semântica e linguística para mineração textual de patentes. **Impacto prático** - a estrutura ontológica viabiliza o uso de ferramentas computacionais para aquisição e organização de inteligência técnica de patentes.
DESENVOLVIMENTO DE UM MÉTODO DE MINERAÇÃO DE INTELIGÊNCIA TÉCNICA EM PATENTES UTILIZANDO UMA ONTOLOGIA BASEADA NOS EFEITOS FÍSICOS DA TRIZ	Desenvolver e validar um método de mineração de inteligência técnica em documentos de patente, utilizando a ontologia como base semântica para aplicações em linguística computacional	Fluxo experimental iniciando com técnicas clássicas de PLN e avançando para arquiteturas sofisticadas, baseadas em LLMs. Extração de documentos de patente da base do INPI por	•Formulário online encaminhado a sete especialistas. •Análise da precisão do método de mineração textual. •Síntese dos instrumentos e/ou ferramentas de coleta e de análise de dados usados no estudo 3	Desenvolvimento de três artigos: •Artigo técnico descrevendo o setup experimental; •Artigo científico descrevendo o método de mineração textual de arquitetura híbrida e utilizando uma base ontológica como recurso de dados semânticos para a extração de termos; •Artigo integrador que consolida os estudos desenvolvidos.	**Produto técnico-tecnológico:** • Método de mineração textual utilizando LLM e Lógica de Decisão Derivativa, disponibilizado em repositório público • Artigo técnico descrevendo o *setup* experimental (em desenvolvimento). **Impacto acadêmico** – desenvolvimento de um método de mineração textual (arquitetura híbrida LLM + Lógica de Decisão) para mineração de patentes em português.

[1] Produtos técnicos-tecnológicos (PTTs) definidos pela CAPES para a Área 27 (Administração Pública e de Empresas, Ciências Contábeis e Turismo).

	e processos de inteligência artificial.	meio de um processo de *crawling*.			**Impacto prático** – fornece uma ferramenta vinculada a uma ontologia que supera as limitações dos modelos exclusivamente supervisionados ou não supervisionados, para a mineração de inteligência técnica em patentes em português. **Impacto social** – indiretamente, apoia o ecossistema de inovação para a resolução dos "grandes desafios sociais" (Wickert et al., 2021). **Impacto político** – o método representa um avanço no fortalecimento do conhecimento técnico e da capacidade de P&D em países de língua portuguesa, influenciando políticas de ciência e tecnologia.

Nota. Adaptado de Costa et al. (2024).

2 PROCEDIMENTOS METODOLÓGICOS

O *Design Science Research*, método de pesquisa adotado na tese, tem seus fundamentos no paradigma das Ciências do *Design*, baseado no trabalho seminal de Herbert Simon, publicado em 1968. Para Simon (1996), a missão de uma Ciência do *Design* é desenvolver conhecimento para a concepção e realização de artefatos, ou seja, para resolver um problema até então não resolvido ou um problema conhecido de uma forma mais eficaz ou eficiente.

O método DSR tem por objetivo gerar conhecimento que seja aplicável e útil para a solução de problemas, melhoria de sistemas já existentes e, ainda, criação de novas soluções e/ou artefatos (Venable, 2006). Portanto, as atividades de pesquisa orientadas para DSR aumentam principalmente o conhecimento aplicável (ou prescritivo) (Hevner et al., 2019; Lacerda et al., 2013), contribuindo para a base do conhecimento.

Os artefatos geram soluções satisfatórias para problemas reais (van Aken, 2004). Podem ser definidos como: constructos, modelos, métodos e instanciações. Os constructos são conceitos que formam o vocabulário do domínio. Os modelos são um conjunto de proposições ou declarações que expressam as relações entre os constructos. Os métodos são um conjunto de etapas para executar uma tarefa (March & Smith, 1995). Por fim, as instanciações são a operacionalização de um artefato ou a articulação de diversos artefatos para a produção de um resultado de um contexto (Lacerda et al., 2013). Mais recentemente, o *framework* é incorporado como um artefato para desenvolver um metamodelo (Sanches et al., 2014).

Os artefatos e suas soluções não constituem uma resposta pontual a um problema específico em determinado contexto (Lacerda et al., 2013). No método DSR, a generalização das prescrições é aplicada a uma classe de problemas (March & Smith, 1995; van Aken, 2004; Venable, 2006), ou seja, a um conjunto de problemas, práticos ou teóricos, que compartilham características comuns (Lacerda et al., 2013).

A estrutura metodológica desta tese fundamenta-se nas seis etapas sequenciais descritas por Peffers et al. (2007), que incluem:

1. Identificação do problema e motivação

São obtidos dados que permitem a compreensão da problemática envolvida (Lacerda et al., 2013) e a justificativa do valor de uma solução (Peffers et al., 2007). Também são verificados os recursos disponíveis e o conhecimento do estado da arte (Dresch et al., 2015).

2. Definição dos objetivos para uma solução de projeto

Os objetivos, quantitativos ou qualitativos, são inferidos a partir da definição do problema e do conhecimento do que é possível e viável. Nesta etapa são também identificadas as soluções atuais e sua eficácia (Dresch et al., 2015).

3. Desenvolvimento do artefato

O artefato é construído para um propósito específico, identificado na etapa 1, e envolve a definição da funcionalidade desejada e da arquitetura correspondente.

4. Demonstração do artefato

Pode incluir experimentação, simulação, estudo de caso, prova ou outra atividade apropriada (Peffers et al., 2007).

5. Avaliação do artefato

Consiste na verificação do desempenho do artefato no ambiente para o qual foi projetado, considerando os objetivos que se propôs a alcançar (March & Smith, 1995). Trata-se de um processo rigoroso (Worren et al., 2002), no qual são comparados os objetivos da solução com os resultados reais observados durante a demonstração do artefato (Peffers et al., 2007). As necessidades de ajustes ou melhorias identificadas nesta etapa são incorporadas no desenvolvimento, em um processo iterativo, ou utilizadas para subsidiar projetos futuros.

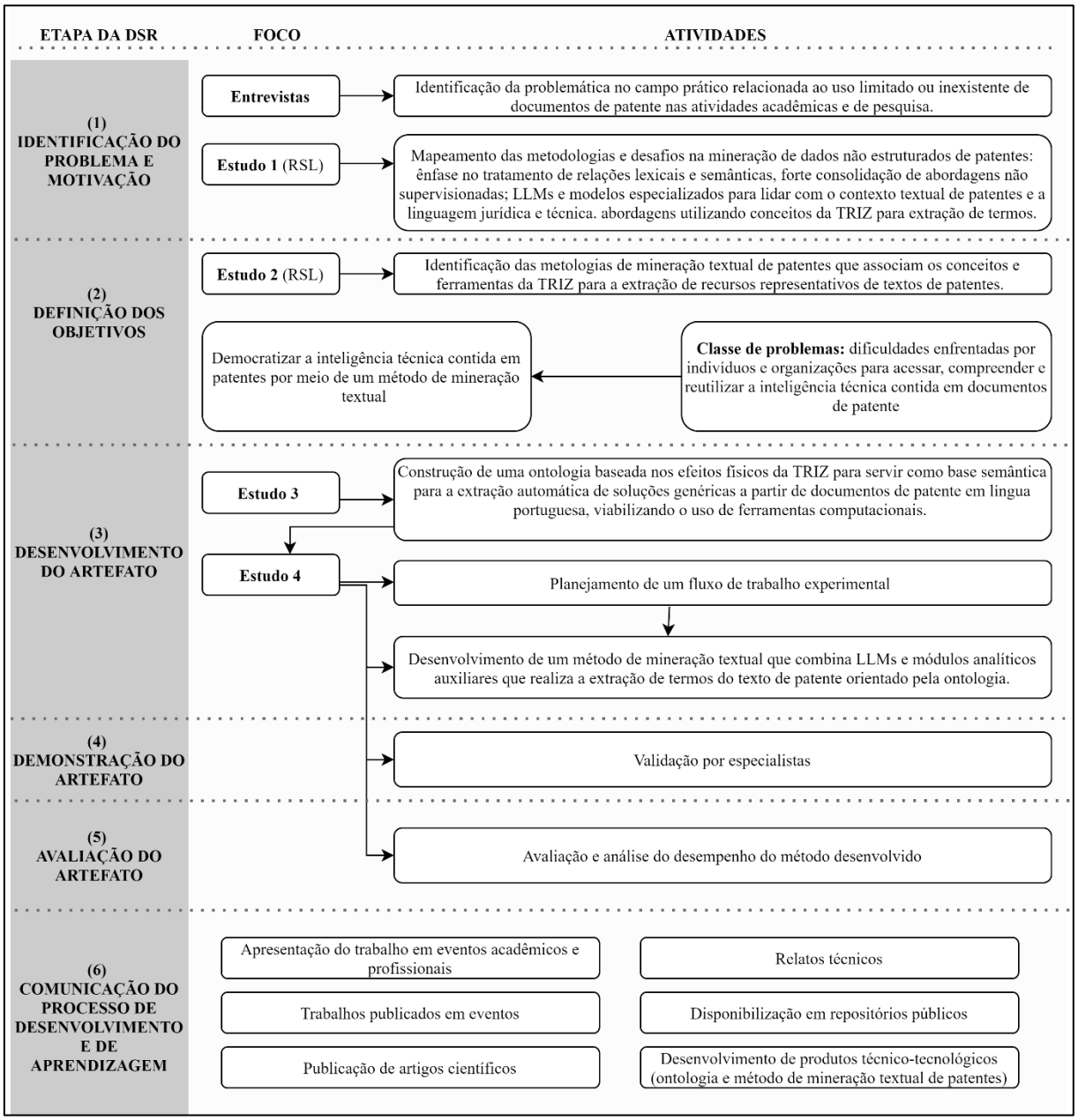
6. Comunicação do processo de desenvolvimento e de aprendizagem

Envolve a difusão do conhecimento gerado (Hevner et al., 2004; Peffers et al., 2007), por meio da apresentação às comunidades acadêmica e profissional (Lacerda et al., 2013). Nesta etapa, são comunicados o problema e sua relevância, a descrição do artefato, sua utilidade e originalidade, bem como o rigor empregado no seu desenvolvimento e avaliação.

As etapas descritas por Peffers et al. (2007) foram sequencialmente organizadas em um diagrama para constituir o desenho da pesquisa. Este desenho metodológico detalha as atividades desenvolvidas em cada etapa, apresentando um resumo dos estudos que compõem esta tese. A Figura 5 ilustra o desenho metodológico da pesquisa, fundamentado nas etapas sequenciais do DSR.

Figura 5

Desenho metodológico da pesquisa com base nas etapas do *Design Science Research*



Nota. Baseado em Peffers et al. (2007).

Nas seções seguintes, serão detalhadamente apresentadas as etapas que constituem o desenho da pesquisa, seguindo a ordem sequencial estabelecida pelo método DSR.

2.1 Etapa 1 - Identificação do Problema e Motivação

Para compreender a problemática no campo prático relacionada à utilização de informações tecnológicas provenientes de patentes nas atividades acadêmicas e de pesquisa,

foram analisados dados obtidos por meio de entrevistas exploratórias com pesquisadores de universidades brasileiras (Cunha et al., 2023). A amostra, definida por conveniência, foi composta por sete entrevistados, com coletas realizadas entre julho e agosto de 2021. Os mapas conceituais gerados a partir das entrevistas evidenciaram a falta de conhecimento e de experiência no uso de bases patentárias, sendo apontadas como principais barreiras a estrutura complexa e a linguagem técnica dessas bases. Embora os participantes reconheçam a relevância das patentes como fonte de informação científica e tecnológica, observou-se uma dependência cultural das bases científicas tradicionais para a construção do conhecimento. A Tabela 2 apresenta as categorias de análise derivadas das entrevistas, acompanhadas de uma breve descrição das informações correspondentes a cada categoria.

Tabela 2

Categorias de análise e trechos das entrevistas

Categoria de análise	Informações obtidas nas entrevistas
Experiência pessoal com bases patentárias	*Eu já olhei alguma coisa do EspaceNet.* *Busca nas bases gratuitas.* *Desconheço.* *Acaba buscando em bases de patentes aqueles alunos que já estão visando fazer uma patente.*
Conhecimento das bases patentárias	*Desconheço* *Bases gratuitas* *EspaceNet* *INPI* *EPO (European Patent Office)*
Motivos para o não uso ou pouco uso das bases patentárias	*Desconhecimento.* *Tradição de utilizar artigos científicos.* *É difícil entender, acessar, utilizar.* *Linguagem um pouco difícil.* *Dificuldade de entender as reivindicações e colocar numa revisão de literatura nem sempre é uma coisa fácil.* *É uma prática bem direcionada, muito restrita a uma parcela de quem faz as pesquisas nas áreas técnico-científicas.*
Percepção sobre a importância do uso das bases de patente	*Aprofundar os estudos.* *Ás vezes a gente acaba recorrendo a bases patentárias para poder suprir uma lacuna que não encontra na literatura convencional.* *Dá um impacto na tua publicação.* *Trazem informações valiosíssimas para a pesquisa.* *Não substitui um outro tipo de pesquisa em artigos científicos, por exemplo, mas complementa.* *Muito relevante, porque ele atalha muito do processo de desenvolvimento, principalmente naquela dimensão mais aplicada, mais entre a interface da pesquisa aplicada com a inovação dentro das empresas.*

Nota. Obtido de Cunha, Volpato e Pedron (2023).

Para compreender a problemática em campo teórico e caracterizar o ambiente de aplicação do artefato a ser proposto, foi realizada uma revisão sistemática da literatura (RSLs) onde foram analisadas as metodologias de mineração textual aplicadas a documentos de patente. Nesse estudo foram identificadas metodologias que empregavam as ferramentas e conceitos da TRIZ na recuperação de inteligência técnica de patentes.

2.2 Etapa 2 - Definição dos Objetivos

Para a identificação das soluções atuais e sua eficácia, nesta etapa da DSR foi realizada uma segunda RSL para aprofundar a compreensão sobre o emprego da TRIZ. Os resultados evidenciaram que as ferramentas da TRIZ oferecem um conjunto de conceitos fundamentais — parâmetros, propriedades e funções — semanticamente alinhados ao vocabulário técnico das patentes, sendo capazes de compor um *corpus* anotado inicial destinado a apoiar aplicações de inteligência artificial voltadas à mineração de patentes.

Considerando o material coletado durante a etapa de identificação do problema, esta segunda fase da DSR teve como foco a definição dos objetivos da solução e a formulação da classe de problemas. A classe de problemas foi estabelecida como as dificuldades enfrentadas por indivíduos e organizações para acessar, compreender e reutilizar a inteligência técnica contida em documentos de patente.

O objetivo da solução foi definido como a democratização do acesso à inteligência técnica contida em patentes, buscando oferecer um artefato que seja, ao mesmo tempo, útil para a prática e relevante para o avanço do conhecimento na área de mineração textual de patentes.

2.3 Etapa 3 - Desenvolvimento do Artefato

Com base no conhecimento derivado dos Estudos 1 e 2, observa-se que as abordagens de mineração textual que empregam as ferramentas da TRIZ apresentam uma singularidade pela proximidade com a linguagem utilizada em patentes, abrangendo termos científicos, tecnológicos e de domínio específico. No entanto, os métodos de mineração textual existentes em sua maioria são desenvolvidos para linguagens de alto recurso, o que os torna pouco adequados ao português, uma linguagem de baixo recurso. As chamadas linguagens de alto recurso são utilizadas em processamento de linguagem natural (PLN) para designar idiomas que dispõem de grande quantidade de dados para treinamento de modelos, além de ferramentas

e recursos linguísticos consolidados, tal como o inglês. As tentativas de adaptação de modelos concebidos para línguas de alto recurso nem sempre se mostram eficazes devido às disparidades linguísticas (Xu et al., 2025), assim como o uso de traduções, que pode negligenciar nuances próprias do idioma e de seus contextos culturais (Majewska et al., 2023).

Além disso, os métodos de mineração textual aplicados para patentes ainda demandam trabalho manual para a rotulação do corpus textual, por meio de métodos supervisionados ou semissupervisionados, para treinar modelos de classificação (Huang & Xie, 2022; Liu, Wu, et al., 2023).

Diante desse cenário, é proposto o desenvolvimento de um método de mineração textual fundamentado em uma ontologia construída a partir de efeitos físicos puros e aplicados. Os efeitos físicos são princípios científicos e fenômenos naturais que, no âmbito da TRIZ, são relacionados em bases de conhecimento para auxiliar nos processos de resolução de problemas (Ilevbare et al., 2013).

Para a construção da ontologia foram extraídos os termos do Banco de Dados Multilíngue de Efeitos Físicos (Zaniro et al., 2024) relacionados a soluções técnicas conceituais, isto é, à definição da função que o produto deve executar e do princípio físico empregado para realizá-la. Em seguida, os termos passaram por um processo de curadoria manual, que incluiu tradução, normalização em subclasses definidas como "Tarefa" (T), "Objeto" (O) e "Efeito Físico" (EF), padronização das classes gramaticais e definição dos relacionamentos entre as subclasses. Esse processo resultou em um relacionamento ternário que integra as subclasses T, O e EF, definido por uma Tarefa (representada por um verbo no infinitivo) que atua sobre um Objeto (representado por um substantivo) e produz um Efeito Físico (definido como um substantivo que expressa um princípio científico ou fenômeno natural). O desenho conceitual da ontologia, juntamente com seus requisitos, constitui o Estudo 3[2].

Para o desenvolvimento do método de mineração textual de documentos de patente, definido como Estudo 4[3], foi estabelecido um *setup* experimental, no qual se construiu um fluxo de trabalho composto por etapas de pré-processamento dos textos brutos (títulos e resumos) de documentos de patente em língua portuguesa obtidos na Base de Patentes do INPI/BR, seguidas

[2] Expresso minha especial gratidão ao Prof. Carlos J. Costa, do Instituto Superior de Economia e Gestão da Universidade de Lisboa, cuja valiosa contribuição foi essencial para a definição da arquitetura da ontologia e para a concepção inicial do método de mineração textual de patentes em português, fundamentado na ontologia como base semântica.

[3] Registro meus sinceros agradecimentos aos colegas do Laboratório de Inteligência Artificial Aplicada da Universidade Federal de Uberlândia, sob a coordenação da Profa. Carla Bonato Marcolin, coorientadora desta tese, pelo apoio e pelas contribuições ao desenvolvimento deste trabalho. Manifesto, em especial, minha gratidão a Patrick Luiz de Araújo e Marcos Antenor de Souza Morais pela colaboração dedicada.

de tratamento semântico. Nesse contexto, foram testadas as seguintes abordagens:

(a) Marcação de classes gramaticais (*Part-of-Speech* – POS): técnica da linguística computacional e do PLN que consiste em atribuir a cada palavra de um texto sua respectiva categoria gramatical (ou classe morfológica).

(b) Comparação de similaridade vetorial: utilizada em PLN para medir o grau de semelhança entre textos, palavras ou documentos, considerando suas representações em forma de vetores numéricos.

(c) Modelos de IA Generativa: aplicados para a geração de novos termos (sinônimos, parônimos e homônimos) a partir de padrões aprendidos nos textos de patentes analisados, permitindo expandir o vocabulário da ontologia usada no treinamento.

Os resultados das três abordagens foram analisados quanto aos termos extraídos e sua proximidade semântica com os resumos e títulos das patentes, bem como com as categorias da ontologia. As duas primeiras abordagens mostraram-se pouco eficientes, ao passo que a última apresentou resultados inicialmente promissores, os quais foram posteriormente submetidos à avaliação de especialistas.

2.4 Etapa 4 - Demonstração do Artefato

Nesta quarta etapa da DSR, a validação dos resultados obtidos por meio do método de mineração foi realizada com a participação de sete especialistas de diferentes áreas do conhecimento, distribuídos geograficamente pelo Brasil. Todos possuíam formação interdisciplinar, contemplando tanto domínio técnico específico nas áreas de Engenharia Mecânica e Elétrica, Farmácia e Química quanto conhecimento aprofundado da matéria de propriedade intelectual e familiaridade com documentos de patente. A seleção dos participantes ocorreu por convite direto, seguido de uma reunião preparatória. Para preservar a imparcialidade e mitigar vieses decorrentes de fatores pessoais ou profissionais, o anonimato dos especialistas foi assegurado por meio de um processo de codificação aplicado aos formulários de resposta e às análises estatísticas, garantindo a proteção da privacidade.

2.5 Etapa 5 - Avaliação do Artefato

Para a avaliação do artefato, foi elaborado um formulário *online* disponibilizado aos

especialistas, contendo, em média, 50 documentos de patente extraídos da base do INPI. O instrumento apresentava quatro questões de resposta única, organizadas em escalas ordinalmente codificadas. O objetivo do formulário era validar os elementos semânticos atribuídos pelo método, verificando sua compatibilidade com o conteúdo do título e do resumo, bem como se o relacionamento semântico ternário (definido por Tarefa–Objeto–Efeito Físico) fornece uma sugestão de solução técnica que a patente se propõe a resolver. Adicionalmente, analisou-se se o relacionamento semântico gerado pelo método apresentava consistência em relação àquele ao qual se vincula na ontologia. Os resultados foram submetidos à análise estatística para o cálculo das taxas de acerto e de erro de consistência, com o propósito de avaliar o desempenho do método de mineração textual e o suporte ontológico na identificação de padrões semânticos e na inferência de relacionamentos técnicos consistentes, com precisão.

2.6 Etapa 6 – Comunicação do Processo de Desenvolvimento

Na etapa de comunicação, os estudos desenvolvidos foram apresentados em eventos científicos e submetidos ou aceitos para publicação, visando à divulgação junto à comunidade acadêmica e profissional. A ontologia, como produto técnico-tecnológico, encontra-se disponível em repositório de acesso público, assim como o *setup* experimental criado para o desenvolvimento do método de mineração textual. A Tabela 3 apresenta os eventos de divulgação dos estudos da tese e os respectivos meios de publicação adotados.

Tabela 3

Síntese dos eventos de divulgação dos estudos e formatos de publicação

Estudo	Evento de divulgação	Meio de publicação
1	XXVI Seminários em Administração da Faculdade de Economia, Administração, Contabilidade e Atuária (FEA) da Universidade de São Paulo (USP) - SEMEAD 2023	Apresentação oral e publicação do resumo Disponível em: https://login.semead.com.br/26semead/anais/resumo.php?cod_trabalho=776
	XXXII Simpósio de Inovação, Tecnologia e Empreendedorismo (SITE 2024) – integralizadas as publicações do ano de 2023 e incluídas as contribuições dos avaliadores do SEMEAD2023	Apresentação oral e publicação do resumo Disponível em: https://anpad.com.br/pt_br/article_search/?search%5Bq%5D=tregnago+cunha&search%5Bsubmit%5D=
	Revista *Management Review Quarterly*	Artigo publicado em setembro de 2025 DOI: 10.1007/s11301-025-00555-

Estudo	Evento de divulgação	Meio de publicação
		z
2	XXVIII Seminários em Administração da Faculdade de Economia, Administração, Contabilidade e Atuária da Universidade de São Paulo (SEMEAD 2025)	Apresentação oral e publicação do resumo Disponível em: https://login.semead.com.br/28semead/anais/resumo.php?cod_trabalho=539
	Revista *World Patent Information*	Submissão de artigo
3	Simpósio Internacional de Gestão, Projetos, Inovação e Sustentabilidade (SINGEP 2024)	Apresentação oral e publicação do resumo Disponível em: https://submissao.singep.org.br/12singep/proceedings/resumo?cod_trabalho=264
	4º *Encuentro de Investigación en Ciencias de la Administración, Universidad Nacional del Sur*/Argentina (2025)	Apresentação oral e publicação do resumo Disponível em: https://repositoriodigital.uns.edu.ar/handle/123456789/7322
	13th World Conference on Information Systems and Technologies (WorldCIST'25)	Apresentação oral
	Emerging Trends in Information Systems and Technologies, Editor Spriger Cham	Artigo publicado em novembro de 2025 DOI: 10.1007/978-3-031-97799-2_19
	Ontologia disponibilizada em repositório público	Produto técnico tecnológico, disponibilizado em repositório público [https://osf.io/cbaef/] (Apêndice A)
4	Método de mineração textual	Produto técnico-tecnológico, disponibilizado em repositório público [https://github.com/PatrickLdA/patent-ai-project] (Apêndice B)
	Artigos em Desenvolvimento: • Artigo técnico descrevendo o *setup* experimental. • Artigo científico descrevendo o método de mineração textual e o desempenho obtido nos experimentos. • Artigo integrador que consolida os estudos desenvolvidos, apresentando uma análise comparativa e conclusiva sobre o desempenho do método e suas aplicações potenciais em mineração de inteligência técnica de patentes.	

Nota: Elaborado pela Autora (2025).

Para assegurar a qualidade linguística do texto, este trabalho utilizou o ChatGPT (OpenAI, 2025) exclusivamente como ferramenta de apoio à revisão gramatical. Ressalta-se que o uso do recurso se limitou à correção e ao aperfeiçoamento da redação, sem qualquer geração de conteúdo original, preservando integralmente a autoria e a integridade das análises apresentadas.

3 ESTUDO 1: REVISÃO SISTEMÁTICA DA LITERATURA SOBRE MINERAÇÃO DE CAMPOS TEXTUAIS DE DOCUMENTOS DE PATENTE

Resumo

Bancos de dados de patentes constituem uma fonte primária de inteligência técnica, ao reunirem informações detalhadas sobre tecnologias recentes e emergentes em diversos domínios. Sob a perspectiva da Visão Baseada no Conhecimento (KBV), que estabelece o conhecimento como o recurso estratégico fundamental das organizações, esses documentos representam ativos de conhecimento codificado cujo valor estratégico depende da capacidade organizacional de reconhecer, assimilar e aplicar o conhecimento neles contido. Nesse contexto, a mineração de texto desempenha papel fundamental ao viabilizar a transformação de grandes volumes de informação técnica dispersa em conhecimento estruturado e acionável. Todavia, esse processo é desafiador em razão do elevado volume de dados, da complexidade estrutural dos textos de patentes e de suas características distintivas, como a combinação de linguagem jurídica e técnica, o caráter multilíngue e os formatos semiestruturados. Este estudo analisa métodos de mineração de texto aplicados a documentos de patentes por meio de uma Revisão Sistemática da Literatura (RSL) de publicações entre 2018 e 2025, indexadas nas bases *Scopus* e *Web of Science*. Ao todo, foram examinados 117 artigos científicos e trabalhos de conferências, possibilitando a identificação de três eixos centrais: (i) tendências na mineração de texto de patentes; (ii) metodologias e ferramentas predominantes para pré-processamento e análise; e (iii) implicações práticas e limitações dos métodos propostos. Os resultados indicam que avanços em processamento de linguagem natural, aprendizado de máquina e aprendizado profundo contribuem para o fortalecimento da capacidade absortiva, ao reduzir barreiras cognitivas e técnicas à aquisição e assimilação do conhecimento externo. Como contribuição prática e gerencial, esta RSL sistematiza avanços metodológicos e recomendações consolidadas, destacando oportunidades de pesquisa futuras. A inteligência técnica extraída de patentes apoia a tomada de decisão estratégica, a inovação e a obtenção de vantagem competitiva, além de contribuir para o avanço dos Objetivos de Desenvolvimento Sustentável (ODS) da Agenda 2030.

Palavras-chave: Patente. Mineração textual. Informação técnica. Visão Baseada no conhecimento.

Abstract

Patent databases constitute a primary source of technical intelligence, as they bring together detailed information on recent and emerging technologies across a wide range of domains. From the perspective of the Knowledge-Based View (KBV), which establishes knowledge as the fundamental strategic resource of organizations, these documents represent codified knowledge assets whose strategic value depends on the organizational capability to recognize, assimilate, and apply the knowledge they contain. In this context, text mining plays a fundamental role by enabling the transformation of large volumes of dispersed technical information into structured and actionable knowledge. However, this process is challenging due to the high volume of data, the structural complexity of patent texts, and their distinctive characteristics, such as the combination of legal and technical language, multilingual content, and semi-structured formats. This study analyzes text mining methods applied to patent documents through a Systematic Literature Review (SLR) of publications from 2018 to 2025 indexed in the Scopus and Web of Science databases. In total, 117 scientific articles and conference papers were examined, allowing the identification of three central axes: (i) trends

in patent text mining; (ii) predominant methodologies and tools for preprocessing and analysis; and (iii) practical implications and limitations of the proposed methods. The results indicate that advances in natural language processing, machine learning, and deep learning contribute to strengthening absorptive capacity by reducing cognitive and technical barriers to the acquisition and assimilation of external knowledge. As a practical and managerial contribution, this SLR systematizes methodological advances and consolidated recommendations, highlighting opportunities for future research. The technical intelligence extracted from patents supports strategic decision-making, innovation, and the achievement of competitive advantage, while also contributing to the advancement of the Sustainable Development Goals (SDGs) of the 2030 Agenda.

Keywords: *Patent. Text mining. Technical information. Knowledge based-view.*

3.1 Introdução

As coleções de documentos de patentes representam uma fonte imensa de conhecimento para comunidades de pesquisa e inovação em todo o mundo (Krestel et al., 2021). A digitalização dos dados de patentes, iniciada em 1984 pelo Instituto Europeu de Patentes (Dintzner & Van Thieleny, 1991), tornou o maior repositório mundial de informações técnicas acessível a um público não especializado (Aristodemou et al., 2017). Atualmente, apenas esse banco de dados oferece acesso a aproximadamente 150 milhões de documentos de patentes multilíngues (European Patent Office, 2025).

Embora as patentes sejam amplamente reconhecidas como uma valiosa fonte de conhecimento científico e técnico, diversos estudos apontam barreiras que dificultam seu uso efetivo pela comunidade acadêmica e por partes interessadas em ambientes industriais e gerenciais nas economias modernas. A terminologia específica de cada domínio, aliada à natureza extensa e heterogênea do conteúdo, acaba por desencorajar seu uso generalizado (Liu, Tan, et al., 2020). Essas características únicas dos textos de patentes, somadas ao grande volume de documentos, exigem o uso de ferramentas especializadas para a recuperação automatizada de informações técnicas (Zhang et al., 2018).

Nas últimas duas décadas, as técnicas de mineração de texto aprimoraram significativamente a capacidade de analisar conjuntos de dados textuais em larga escala. No domínio das patentes, os avanços no processamento e na análise automatizados (Antons et al., 2020) facilitaram a extração mais eficaz de informações relevantes a partir de dados não estruturados (Krestel, Chikkamath, et al., 2021; Krestel et al., 2023). No entanto, patentes não são inerentemente legíveis por máquinas, exigindo, com frequência, uma análise cuidadosa e intervenção humana (Prickett & Aparicio, 2012). As limitações na vinculação precisa entre

tecnologias-chave e suas respectivas funções comprometem a interpretabilidade e a aplicabilidade prática dos resultados de pesquisa (Giordano et al., 2023).

Um número ainda relativamente pequeno de revisões sistemáticas da literatura reúne as metodologias empregadas na mineração de texto de patentes, permitindo identificar tendências e potenciais áreas de investigação, além de servir como um recurso valioso para pesquisadores, como Antons et al. (2020), Chuprat et al. (2024) e An et al. (2024), este último com foco na síntese de ferramentas computacionais avançadas para a análise de patentes. Essas contribuições representam avanços acadêmicos significativos, apresentando abordagens diversas que refletem os desenvolvimentos, metodologias e aplicações mais recentes na área.

No entanto, em estudos empíricos, os *insights* mais valiosos frequentemente emergem das discussões sobre limitações e direções futuras, e não apenas das descrições metodológicas. Sistematizar esse tipo de conhecimento pode contribuir para a identificação de lacunas na pesquisa, orientar investigações subsequentes e assegurar que estudos futuros abordem deficiências já reconhecidas (Montgomery, 2023).

Nesse contexto, a contribuição deste estudo está na identificação das metodologias aplicadas à mineração de dados não estruturados de patentes, nos principais desafios enfrentados ao longo do processo analítico e nas recomendações oferecidas pelos pesquisadores, com o objetivo de informar e orientar futuras investigações.

Este trabalho contribui para o corpo de conhecimento existente ao abordar a seguinte questão de pesquisa: Quais são as contribuições das metodologias de mineração de textos de patentes para orientar estudos subsequentes? Seu objetivo principal é mapear e avaliar metodologias de mineração de textos aplicadas a documentos de patentes, examinando a evolução dessas metodologias, suas limitações e os caminhos potenciais para pesquisas futuras.

Como contribuição central, este estudo consolida o conhecimento atual e as metodologias já estabelecidas no campo da mineração de textos de patentes, ao mesmo tempo em que identifica lacunas, evidencia a necessidade de aprimoramentos e propõe direções promissoras para futuras investigações. A mineração de textos emerge, nesse contexto, como um mecanismo essencial para a transformação da informação técnica codificada em conhecimento organizacional estratégico, ao fornecer inteligência para o desenvolvimento competitivo de produtos e serviços, a identificação de novas aplicações tecnológicas e a antecipação de tendências emergentes, frequentemente antes mesmo de seu reconhecimento na literatura científica. Entre suas implicações práticas e gerenciais, destaca-se ainda o papel da mineração de textos como uma ferramenta poderosa de apoio aos ODS, ao possibilitar a análise de grandes volumes de dados e a identificação de *insights* relevantes para a tomada de decisão.

O estudo está organizado em seções principais**.** A Introdução apresenta o contexto e os objetivos da pesquisa. A Revisão da Literatura discute as principais abordagens e aplicações da mineração de texto direcionadas à extração de informações de patentes. A seção de Metodologia descreve detalhadamente o processo de estruturação da Revisão Sistemática da Literatura (RSL). A seção de Resultados está subdividida em três partes: a primeira analisa o perfil das publicações selecionadas, considerando a nacionalidade dos autores, as redes de colaboração, os periódicos e as áreas de publicação; a segunda examina as metodologias e ferramentas empregadas nos estudos; e a terceira sintetiza as contribuições práticas e as recomendações para pesquisas futuras identificadas nos artigos revisados. Por fim, a seção de Discussão e Comentários Finais aborda as contribuições teóricas do estudo, suas implicações práticas e gerenciais, bem como suas limitações e sugestões para investigações futuras.

3.2 Revisão da Literatura

Na revisão da literatura, são discutidas as características peculiares dos documentos de patente em relação à sua estrutura e linguagem, aspectos que os diferenciam dos documentos científicos. A segunda subseção trata dos processos de mineração textual aplicados a patentes, destacando as principais abordagens utilizadas para a extração de termos, bem como os modelos de dados empregados na identificação de informações tecnológicas relevantes.

3.2.1 Características dos Documentos de Patente: Estrutura e Linguagem

As patentes são documentos técnicos especializados, caracterizados por um conjunto padronizado e consistente de elementos textuais. Representam, segundo Shelick (2009, p. 53), "o resultado da intersecção de diferentes disciplinas, expresso em uma materialidade linguística normativa e descritiva". Nesse sentido, são amplamente reconhecidas como uma das fontes mais confiáveis de inteligência técnica (Lee et al., 2019), desempenhando papel central tanto na análise do progresso científico e tecnológico (Pilkington et al., 2009; Zhang et al., 2011) quanto no apoio à previsão tecnológica.

O conhecimento técnico disponível em patentes apresenta linguagem e estrutura próprias (Sun et al., 2021). Esses documentos combinam dados estruturados e não estruturados. Os dados estruturados incluem metadados, como titular da patente, inventores e códigos de classificação que vinculam a invenção a domínios tecnológicos específicos (Krestel et al., 2021; Sun et al., 2021). Já os dados não estruturados abrangem elementos como título, resumo,

descrição detalhada da invenção, reivindicações e figuras associadas, componentes que não seguem um formato rígido (Khadilkar et al., 2019).

Nesse contexto, os classificadores de patentes assumem papel fundamental, pois são sistemas utilizados para organizar, indexar e recuperar documentos em bases de dados. Entre os principais classificadores adotados internacionalmente, destacam-se:

(a) Classificação Internacional de Patentes (IPC – *International Patent Classification*), criada pela Organização Mundial da Propriedade Intelectual (WIPO – *World Intellectual Property Organization*), composta por oito seções principais subdivididas em classes, subclasses, grupos e subgrupos (World Intellectual Property Organization, 2025a);

(b) Classificação Cooperativa de Patentes (CPC - *Cooperative Patent Classification*), desenvolvida em conjunto pelo Escritório Europeu de Patentes (EPO) e pelo Escritório de Patentes dos Estados Unidos (USPTO), baseada na IPC, porém com maior nível de detalhamento (European Patent Office, 2025b).

Por outro lado, quando se trata dos dados não estruturados, as tecnologias de mineração de texto desempenham papel essencial. Essas ferramentas permitem extrair e processar o conteúdo textual, convertendo-o em informações semânticas relevantes por meio de diferentes técnicas analíticas (Ki & Kim, 2017). No contexto das patentes, a mineração de texto tem ganhado destaque devido ao crescimento exponencial do volume de documentos, que impõe desafios consideráveis ao acesso eficiente às informações mais recentes e relevantes, especialmente para profissionais envolvidos em processos de patenteamento, bem como para leitores não especializados ou sem formação técnica (Chiarello et al. 2018).

No entanto, a linguagem das patentes é altamente terminológica, marcada pelo uso de termos científicos, jargões jurídicos e sinônimos (Hu et al., 2018; Kim, Choi, et al., 2018; Xie & Miyazaki, 2013). Além disso, apresenta frases longas e complexas, com estrutura muitas vezes propositalmente ambígua (Berdyugina & Cavallucci, 2021). Em razão dessas particularidades de expressão e da especificidade da terminologia (Yue, Liu, Zhang, et al., 2023), a mineração de textos de patentes é dificultada pela presença de expressões redundantes, pela polissemia, isto é, a atribuição de múltiplos significados em diferentes contextos, e pela complexidade inerente ao conteúdo interdisciplinar (Hu et al., 2018).

3.2.2 Mineração Textual de Documentos de Patente

A mineração de textos de patentes pode ser classificada em quatro estratégias principais: (i) gestão do conhecimento ou conhecimento de domínio (Ghoula et al., 2007; Kim et al., 2018;

Lee et al., 2014); (ii) mapeamento de patentes (Lei et al., 2019); (iii) gestão tecnológica; e (iv) extração de informações (Gerken & Moehrle, 2012).

A estratégia baseada em conhecimento de domínio concentra-se na avaliação da qualidade das patentes e no apoio a tarefas de classificação, frequentemente por meio da construção de ontologias específicas, que possibilitam a categorização e recuperação automáticas de documentos (Aristodemou & Tietze, 2018; Lee et al., 2019). O mapeamento de patentes, por sua vez, visa à visualização de patentes em um espaço bidimensional com base na similaridade de seus conteúdos tecnológicos (Lin et al., 2022; Wang & Chen, 2019). A gestão tecnológica foca no monitoramento da evolução de tecnologias e na previsão de inovações emergentes (Aristodemou & Tietze, 2018). Por fim, a extração de informações busca identificar automaticamente elementos específicos como nomes químicos, dados numéricos e trechos descritivos relevantes (Aristodemou & Tietze, 2018; Sarica et al., 2020).

Em qualquer uma das estratégias, três abordagens principais são comumente empregadas para a extração de termos:

(a) Métodos supervisionados dependem de *corpora* rotulados para treinar modelos de classificação ou utilizam bases de conhecimento externas como suporte (Huang & Xie, 2022; Liu, Wu, et al., 2023);

(b) Métodos não supervisionados "focam na mineração de conexões internas em documentos em resposta à falta de dados rotulados" (Liu, Wu, et al., 2023, p. 129:4);

(c) Métodos semissupervisionados combinam dados rotulados e não rotulados, rotulando progressivamente estes últimos com base nos classificadores iniciais. Esse processo iterativo permite treinar simultaneamente modelos com ambos os tipos de dados, aumentando a precisão da classificação (Rosenberg et al., 2005; Triguero et al., 2015).

A Tabela 4 apresenta a síntese dos métodos de extração de palavras-chave aplicados nas estratégias de mineração de texto, destacando seus respectivos pontos fortes e limitações.

Tabela 4

Síntese dos métodos de extração de palavras-chave em estratégias de mineração de texto e respectivos pontos fortes e fracos

Métodos de extração de palavras-chave	Pontes fortes	Pontos fracos
Extração supervisionada	Esse método geralmente se baseia na frequência de ocorrência de um termo	Esse método exige um corpus rotulado, cuja qualidade das

Métodos de extração de palavras-chave	Pontes fortes	Pontos fracos
	e em sua posição no documento durante o processo de classificação. (Aggarwal et al., 2018).	anotações influencia diretamente o desempenho do modelo (Hu et al., 2018). No entanto, no contexto da literatura de patentes, ainda não existe um corpus anotado direcionado, abrangente e em larga escala (Liwei, 2022). Como consequência, a rotulagem manual torna-se necessária para treinar funções de classificação, um processo custoso, demorado e suscetível a erros humanos (Hu et al., 2018). Além disso, esses métodos requerem grandes volumes de dados de treinamento para alcançar resultados eficazes (Dessi et al., 2023), sendo que o desempenho da extração de termos depende tanto da qualidade quanto da escala do *corpus* utilizado (Li et al., 2017).
Extração não supervisionada	Vários critérios de pontuação foram desenvolvidos para a classificação de palavras-chave candidatas, incluindo análise linguística, métodos estatísticos, abordagens de modelagem de tópicos e técnicas baseadas em grafos de rede (Huang & Xie, 2022; Ren et al., 2022). Uma vantagem fundamental desses métodos é que eles não exigem um corpus rotulado manualmente, tornando-os particularmente adequados para cenários com dados anotados limitados (Hu et al., 2018).	A coleta de um corpus específico de domínio continua sendo uma tarefa desafiadora (Florescu and Caragea, 2017). lém disso, métodos não supervisionados frequentemente enfrentam dificuldades para extrair palavras-chave semanticamente significativas (Dessi et al., 2023), o que pode comprometer a precisão da extração (Joshi et al., 2022).
Extração semi-supervisionada	Esse método requerer uma pequena quantidade de corpus anotado para extrair todos os relacionamentos dentro de um documento (Wang et al., 2021).	Esse método requer a construção de um dicionário que contém palavras candidatas (palavras-semente) que potencialmente definem as categorias relevantes (Watanabe & Zhou, 2022). No entanto, a presença de palavras-semente correspondentes a textos não relacionados pode degradar significativamente o desempenho do classificador, pois falsos positivos obscurecem as

Métodos de extração de palavras-chave	Pontes fortes	Pontos fracos
		verdadeiras associações entre tópicos e palavras durante o processo de aprendizado de máquina (Watanabe & Zhou, 2022).

Nota: Dados da pesquisa (2025).

No entanto, a extração de termos específicos de domínio na literatura de patentes permanece uma tarefa desafiadora (Liwei, 2022). Os campos não estruturados desses documentos frequentemente apresentam terminologia que reflete as características gerais da linguagem científica e tecnológica. De acordo com os princípios fundamentais dessa terminologia, termos-chave (simples ou compostos) tendem a estar associados a domínios tecnológicos específicos (Liwei, 2022). Contudo, devido à significativa variação no conteúdo técnico entre diferentes áreas, certos termos aparecem com elevada frequência em um domínio, mas são raramente empregados em outros (Liwei, 2022; Wang et al., 2019). Além disso, funções semelhantes em domínios distintos costumam ser descritas por terminologias técnicas completamente distintas, o que torna ainda mais complexa a compreensão e a análise interdomínios (Arts et al., 2021).

Dentre os diversos modelos de dados utilizados para extrair informações tecnológicas relevantes, a extração de palavras-chave é amplamente reconhecida na literatura como uma etapa fundamental para tarefas analíticas subsequentes (Hu et al., 2018; Xie & Miyazaki, 2013; Yoon & Magee, 2018). Essa técnica serve de base para a identificação de tendências, relacionamentos e inovações em dados de patentes (Kim et al., 2018). No entanto, abordagens tradicionais de mineração de texto que dependem fortemente da extração de palavras-chave apresentam diversas limitações. Entre elas, destacam-se a dificuldade em capturar o significado contextual dos termos, a possível omissão de palavras-chave críticas e a tendência a focar principalmente em substantivos e verbos, negligenciando, assim, outras classes gramaticais relevantes (Joshi et al., 2022). Como observam Kim, Choi, et al. (2018, p. 534), “uma única palavra é muito fragmentária para fornecer informações tecnológicas”.

Para superar a perda de relações semânticas durante a extração de termos (Aristodemou and Tietze, 2018), têm sido propostos modelos linguísticos mais avançados. Entre eles, destacam-se as abordagens baseadas nas estruturas SAO (Jang et al., 2022; Yoon et al., 2013) e no modelo Propriedade-Função (PF) (Kim et al., 2018), ambas com o objetivo de capturar informações contextuais e relacionais mais ricas. A Tabela 5 sumariza as abordagens de

extração de termos aplicadas a patentes.

Tabela 5

Síntese das abordagens para extração de termos de textos de patentes e respectivos pontos fortes e fracos

Abordagem de extração de termos de patente	Pontes fortes	Pontos fracos
Seleção de palavras-chave	Essa abordagem considera a distribuição de palavras-chave em documentos de patentes, levando em conta métricas como a frequência de termos e a frequência de termos ponderada pela frequência inversa nos documentos (TF-IDF). Essa abordagem avalia a relevância dos termos principalmente com base em sua ocorrência tanto dentro de um documento específico quanto em relação ao conjunto de documentos analisados (Lee et al., 2009; Noh et al., 2015).	ssa abordagem não considera as características semânticas do texto (Salton & Buckley, 1988) e desconsidera a influência da posição da frase dentro do documento de patente (Papagiannopoulou & Tsoumakas, 2020). Além disso, de forma convencional, não incorpora termos compostos por duas ou mais palavras, como bigramas e trigramas.
SAO	Em textos de patentes, os objetos, ferramentas, métodos e sistemas descritos em uma invenção podem frequentemente ser representados como o Assunto (S) em frases que detalham a tecnologia subjacente (Park et al., 2013). A função, definida como a tarefa ou ação que um sistema ou tecnologia executa, pode normalmente ser expressa na forma de uma estrutura Ação-Objeto (AO) (Park et al., 2013).	A estrutura SAO é limitada em sua capacidade de analisar informações textuais contendo verbos que não foram previamente definidos (Jang & Yoon, 2021). Além disso, o sujeito e o objeto não possuem tipos de entidade explícitos e a ação não transmite uma relação semântica específica entre eles. (Chen et al., 2022).
Propriedade-Função (PF)	As propriedades são descritas principalmente por meio de adjetivos, enquanto as funções costumam ser expressas por verbos (Yoon, 2010). Essa distinção é particularmente relevante no contexto tecnológico, uma vez que propriedades frequentemente dão origem a funções (Dewulf, 2011). Além disso, pares de palavras do tipo propriedade–função oferecem um nível adequado de complexidade para a construção de redes (Kim et al., 2018).	A utilização apenas de adjetivos e verbos não representa suficientemente concretude dos domínios tecnológicos (Yoon & Kim, 2012).

Nota: Dados da pesquisa (2025).

Apesar dos avanços significativos nos processos de extração de termos, a extração automática de palavras-chave ainda enfrenta diversos desafios. Entre os principais, destacam-se a ocorrência de expressões redundantes, a polissemia (palavras que assumem múltiplos significados conforme o contexto), a necessidade de atualizações contínuas nos tesauros, a complexidade do conteúdo interdisciplinar e o elevado grau de intervenção manual requerido (Hu et al., 2018; Krestel, Chikkamath, et al., 2021; Liu et al., 2021). O desenvolvimento de tecnologias de extração de informação tem, em geral, seguido três etapas principais: métodos baseados em regras, métodos fundamentados em aprendizado de máquina e, mais recentemente, abordagens que empregam aprendizado profundo (*Deep Learning* – DL) (Yang et al., 2022). A Tabela 6 apresenta a síntese dos métodos de extração de informações de textos, destacando suas principais características e potenciais aplicações.

Tabela 6

Síntese dos métodos de extração de informações, características e potenciais aplicações

Método de extração de informações	Características	Potenciais aplicações
Método baseado em regras	Esse método envolve a construção manual de regras linguísticas para representar as características semânticas das frases. É utilizado tanto no reconhecimento de entidades quanto na identificação e classificação de relacionamentos entre elas (Yang et al., 2022).	Essa abordagem pode alcançar alta precisão em pequenos conjuntos de dados, mas apresenta limitações de escalabilidade e adaptabilidade quando aplicada a bases maiores ou mais heterogêneas (Yang et al., 2022). Além disso, exige tempo e esforço humano significativos para a construção e manutenção dos conjuntos de regras, devendo ser adaptada manualmente a diferentes idiomas, o que restringe sua eficiência e capacidade de generalização.
Método baseado em aprendizado de máquina	Esse método utiliza treinamento supervisionado a partir de um corpus rotulado manualmente. Em seguida, o modelo treinado é aplicado para classificar ou extrair informações relevantes de novos dados (Miric et al., 2023; Yang et al., 2022).	Esta abordagem requer anotação manual por indivíduos com experiência específica no domínio, tornando o processo trabalhoso e demorado (Yang et al., 2022).
Método baseado em aprendizado profundo	Identifica automaticamente padrões de informação e estruturas subjacentes por meio do uso de arquiteturas de redes neurais complexas (Yang et al.,	Este método é particularmente adequado para processamento de *big data*, pois pode aprender automaticamente características de frases sem a necessidade de engenharia de recursos manual ou complexa (Yang et

Método de extração de informações	Características	Potenciais aplicações
	2022).	al., 2022).

Nota: Dados da pesquisa (2025).

No domínio da IA, dados textuais não estruturados vêm sendo progressivamente transformados em representações formais e legíveis por máquinas, a fim de viabilizar o processamento e o aprendizado automatizados (Jiang & Shang, 2020; Krestel, Chikkamath, et al., 2021; Singh, 2018; Trappey, Trappey, & Chang, 2020; Yoon & Park, 2005). Esses métodos facilitam a extração de informações relacionais entre unidades semânticas (Sarica et al., 2020), automatizando tarefas que antes demandavam *expertise* especializada (Krestel et al., 2021). As técnicas evoluíram significativamente na extração de *insights* a partir de *big data* (Abbas et al., 2014; Chen et al., 2020; Lupu, 2017; Oldham & Fried, 2016; Trippe, 2015). Como resultado, a aplicação de métodos de IA, incluindo aprendizado de máquina (*Machine Learning* - ML), redes neurais artificiais de modelos de linguagem ampla (*Largue Language Model* - LLM) e DL, tem ganhado atenção crescente no campo da propriedade intelectual, refletida no aumento contínuo do número de publicações e citações (Aristodemou & Tietze, 2018; Blume et al., 2024; Jiang et al., 2025; Zhang et al., 2023).

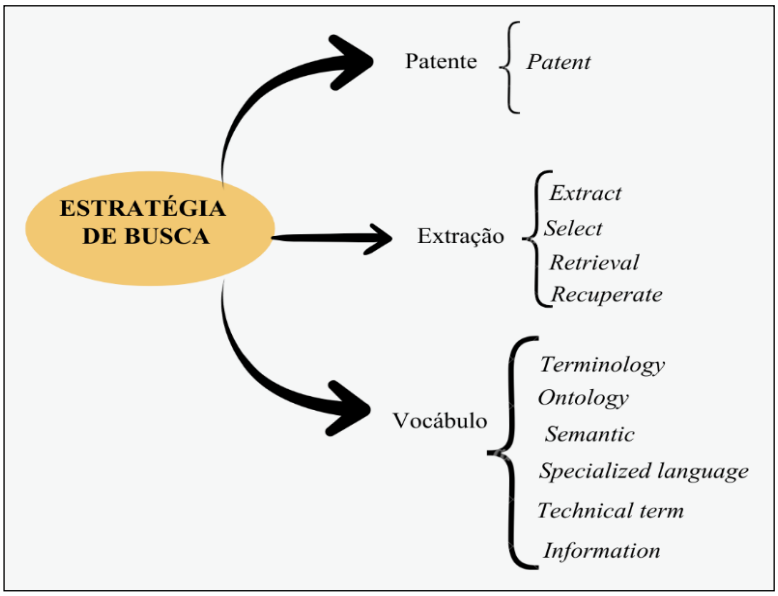
3.3 Procedimentos Metodológicos

Para investigar o corpo de conhecimento sobre técnicas de mineração de texto aplicadas à extração de informações em textos de patentes, foi conduzida uma Revisão Sistemática da Literatura (RSL), seguindo protocolos estabelecidos para o gerenciamento de grandes corpora documentais (Galvão and Ricarte 2019). Essa revisão adota as diretrizes propostas no *Cochrane Handbook for Systematic Reviews of Interventions (2023)*, que oferece uma estrutura metodológica para a definição de critérios de inclusão e exclusão.

Na etapa de seleção das publicações, foram utilizadas as bases de dados *Scopus* e *Web of Science*, escolhidas por seu escopo interdisciplinar e ampla cobertura da literatura científica em diversas áreas do conhecimento. Para a definição dos termos de busca empregados no processo de recuperação dos artigos, realizou-se um estudo piloto fundamentado em manuais da área de terminologia, complementado pela consulta a dicionários de sinônimos. O mapa conceitual que orienta esse processo encontra-se apresentado na Figura 6.

Figura 6

Mapa conceitual de termos para a estratégia de busca para a Revisão Sistemática da Literatura do Estudo 1



Nota. Elaborado pela Autora (2025).

Os termos presentes no mapa conceitual foram conectados por operadores booleanos, resultando na seguinte estratégia de busca para a recuperação de documentos: [*patent AND (extract* OR select OR retrieval OR recuperate) AND (terminology OR ontology OR semantic OR "specialized language" OR "technical term" OR information)*]. A busca foi aplicada aos campos de título, resumo e palavras-chave em ambas as bases de dados. Foram considerados artigos de periódicos e de conferências, publicados entre 2018 e maio de 2025, nos idiomas inglês, espanhol e português.

Para conduzir a RSL foi adotada a estrutura PRISMA (*Preferred Reporting Items for Systematic Reviews and Meta-Analyses*) (PRISMA 2025). Essa escolha foi fundamentada em dois aspectos principais: (i) sua abordagem rigorosa e sistemática para definir critérios de inclusão e exclusão e implementar estratégias de busca; e (ii) sua capacidade de gerar um diagrama de fluxo que mapeia de forma clara o número de registros identificados, incluídos e excluídos, juntamente com os motivos de exclusão, aumentando, assim, a transparência e a interpretabilidade dos resultados (Page et al., 2021).

Na fase inicial da RSL, denominada fase de identificação, foram recuperados 1.739 artigos, posteriormente importados para a plataforma Rayyan (Ouzzani et al., 2016). Nessa etapa, 501 artigos duplicados foram removidos.

Durante a fase de triagem, os títulos e resumos de 1.238 artigos foram analisados, resultando na exclusão de 618 publicações não pertinentes ao escopo do estudo. As 620

publicações remanescentes passaram à avaliação em texto completo, da qual emergiram 117 artigos incluídos na revisão final. As 620 publicações remanescentes passaram à avaliação em texto completo, da qual emergiram 117 artigos incluídos na revisão final.

Entre os 503 artigos excluídos nessa fase:

(a) 46,6% (n=289) abordavam métodos como sumarização, classificação, tradução, análises patentométricas ou estratégias de busca;

(b) 19,8% (n=123) tratavam de metodologias de mineração de texto aplicadas a línguas não ocidentais (por exemplo, línguas orientais ou russo) ou careciam de descrição metodológica;

(c) 7,9% (n=49) enfocavam a mineração de estruturas químicas ou biológicas em documentos de patentes;

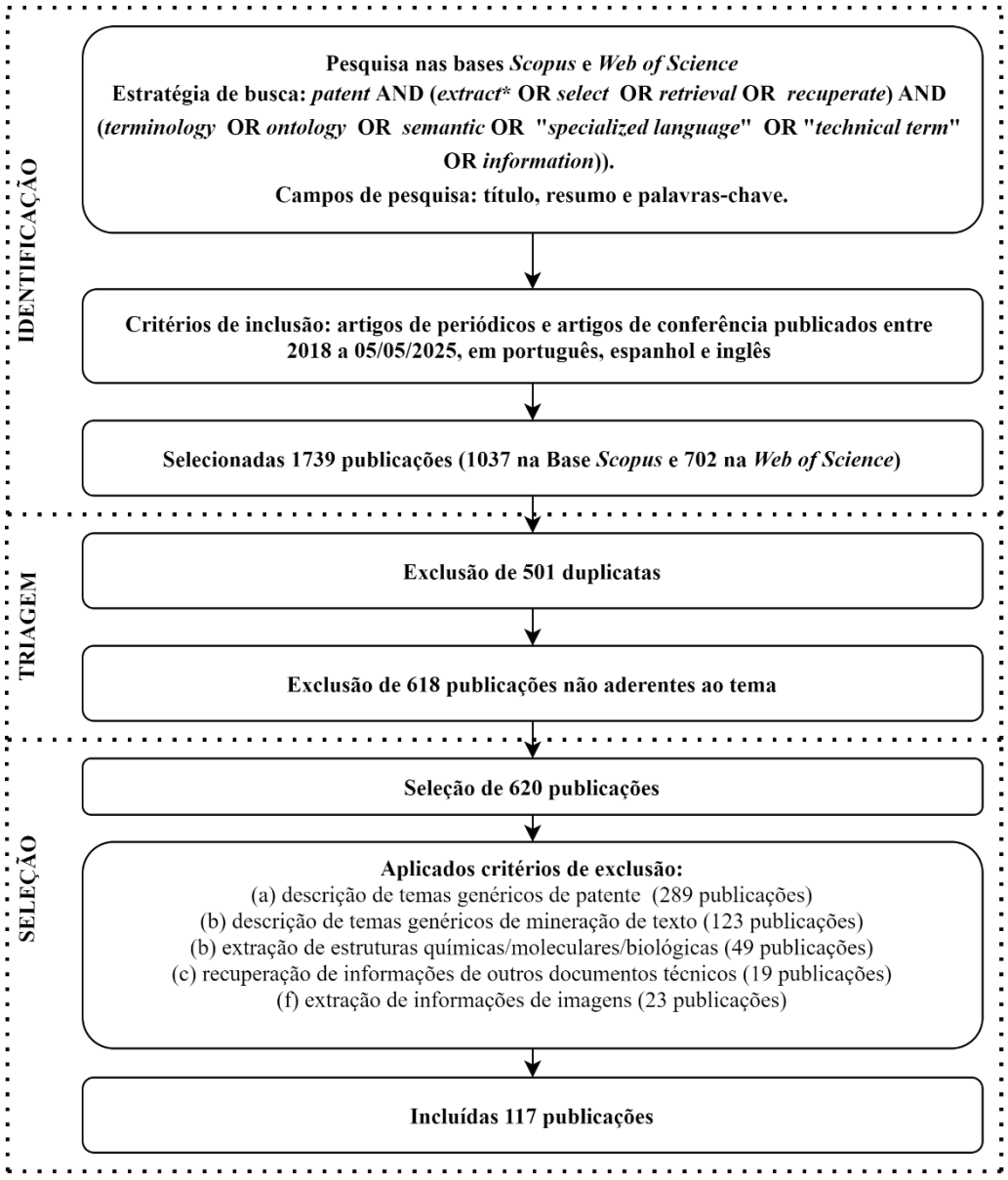
(d) 3,7% (n=23) exploravam a extração de informações a partir de imagens;

(e) 3,0% (n=19) envolviam a mineração de texto em documentos técnicos não relacionados a patentes.

A sequência metodológica adotada na pesquisa é consolidada na Figura 7, que ilustra as etapas de refinamento e seleção do *corpus* bibliográfico.

Figura 7

Diagrama de fluxo da Revisão Sistemática da Literatura do Estudo 1



Nota. Adaptado de PRISMA (2025).

Após a revisão das 117 publicações incluídas na RSL (Apêndice C), foram analisadas informações bibliográficas com o objetivo de caracterizar o perfil dos estudos. Além disso, foram examinados aspectos técnicos, com ênfase nas metodologias e aplicações descritas, nos métodos de extração de termos técnicos adotados, nas abordagens de extração empregadas, nos domínios específicos de patentes abordados e nas vantagens e desvantagens relatadas das ferramentas utilizadas.

3.4 Resultados

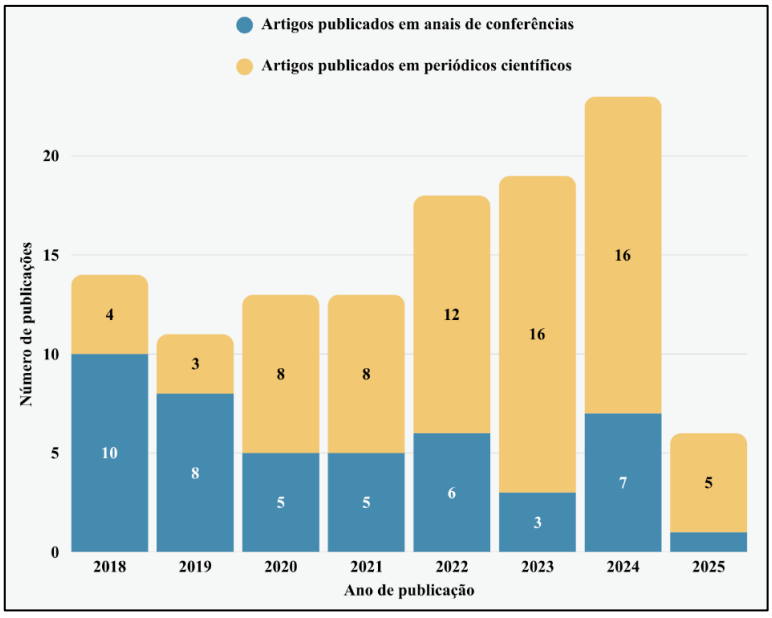
A apresentação dos resultados da RSL está organizada em duas seções. A primeira contempla a análise das informações bibliográficas das publicações selecionadas, com o objetivo de caracterizar o perfil da área de estudo. A segunda concentra-se nos dados técnicos, destacando as estratégias de mineração de informação, os métodos de extração de termos, as técnicas de análise de dados textuais e os campos de patentes mais explorados.

3.4.1 Análise das Informações Bibliográficas das Publicações Selecionadas na Revisão Sistemática da Literatura

A análise das informações bibliográficas obtidas na RSL, referente ao período de 2018 a parte de 2025, revela que 59% das publicações foram veiculadas em periódicos, enquanto os 41% restantes foram apresentadas em congressos. No intervalo considerado, o número total de publicações manteve-se relativamente estável, com um crescimento mais perceptível a partir de 2022. Observa-se, ainda, uma mudança nas tendências de disseminação, marcada por uma transição gradual das apresentações em congressos para publicações em periódicos. A Figura 8 apresenta o gráfico da distribuição das publicações de acordo com o meio de divulgação, no período de 2018 a 2025.

Figura 8

Gráfico da distribuição das publicações científicas selecionadas na Revisão Sistemática da Literatura do Estudo 1, de acordo com o meio de divulgação



Nota. Dados da pesquisa (2025).

Entre os artigos selecionados, em sua maioria oriundos das áreas de Ciência da Computação e Engenharia, observou-se que os trabalhos apresentados em conferências geralmente trazem resultados preliminares, com descrições sucintas das metodologias e análises de dados limitadas. Em contraste, os artigos publicados em periódicos tendem a apresentar revisões bibliográficas mais abrangentes, metodologias detalhadas e análises completas dos resultados.

Verificou-se também que os artigos de conferência costumam ter um número maior de autores, com média de três por publicação, além de frequentemente receberem citações recorrentes ao longo do tempo. O declínio observado na quantidade de publicações em conferências ao longo do período analisado pode indicar um processo de amadurecimento da pesquisa, culminando em resultados mais consolidados, apropriados para discussões aprofundadas em periódicos.

No entanto, alguns grupos de pesquisa, particularmente os afiliados à Universidade de Volgograd e ao Instituto Nacional de Ciências Aplicadas de Estrasburgo, priorizam quase exclusivamente publicações em conferências.

A análise das afiliações dos artigos selecionados na RSL revela que a China concentra o maior número de autores contribuintes, seguida pela Coreia do Sul. De forma agregada, os países asiáticos dominam o cenário global, respondendo por 71,75% de todos os autores. Os países europeus representam 24,63%, enquanto as Américas, sobretudo Estados Unidos e Canadá, contribuem com 3,4%. Já os autores africanos correspondem a apenas 0,22%.

Uma possível explicação para a predominância asiática é o aumento expressivo no número de pedidos de patentes na China e na Coreia do Sul, economias líderes em inovação no Sudeste, Leste Asiático e Oceania — sendo a China amplamente reconhecida como uma das potências globais de inovação (World Intellectual Property Organization, 2024).

A Tabela 7 apresenta a frequência das afiliações institucionais das publicações selecionadas na RSL, listadas em ordem decrescente.

Tabela 7

Frequência das afiliações institucionais das publicações selecionadas na Revisão Sistemática da Literatura do Estudo 1

Afiliação	Número de autores [a]	Afiliação	Número de autores [a]
China	217	Indonésia	05
Coréia	47	Canadá	04
Itália	27	Países Baixos	04

Afiliação	Número de autores [(a)]	Afiliação	Número de autores [(a)]
Federação Russa	24	Brunei	03
Taiwan	23	Vietnã	03
França	21	Finlândia	02
Alemanha	19	Reino Unido	02
Estados Unidos	11	Dinamarca	01
Singapura	08	Paquistão	01
Malásia	07	Maurício	01
Áustria	06	Índia	01

Nota. Dados da pesquisa (2025).
[(a)] Artigos com mais de uma afiliação são multicomputados.

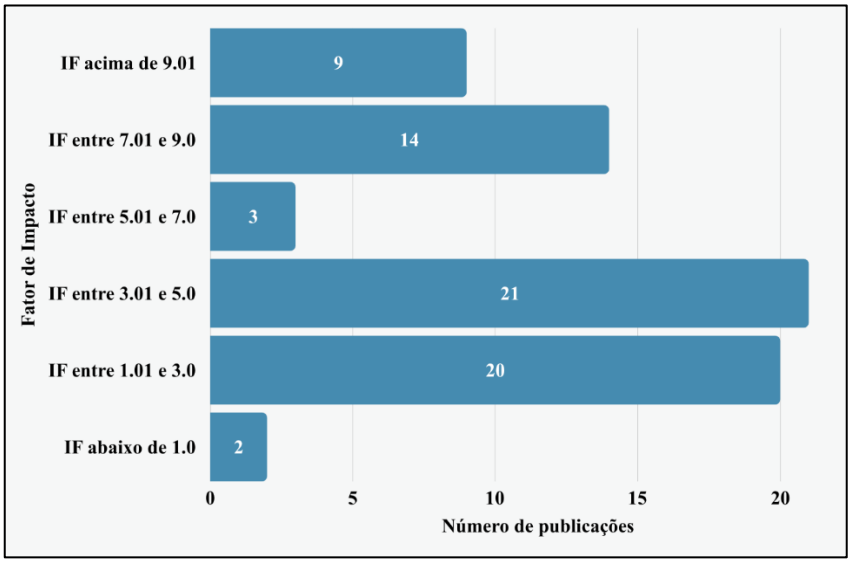
Pouco mais de 10% das publicações selecionadas relatam resultados de pesquisas transnacionais. Refletindo sua forte presença no conjunto de dados, autores chineses aparecem com frequência em colaborações com pesquisadores dos Estados Unidos, Canadá e Europa. Contudo, muitas dessas iniciativas internacionais parecem originar-se de projetos acadêmicos liderados por pesquisadores emergentes, que não voltam a aparecer em publicações subsequentes.

A partir de 2022, observou-se um crescimento significativo nas pesquisas colaborativas, com o número de publicações desse tipo dobrando em comparação ao período de 2018 a 2021. Essa tendência pode indicar uma necessidade crescente de fortalecer a produção científica nacional por meio da cooperação internacional, favorecendo a troca de conhecimento e a busca por soluções mais robustas.

Para compreender melhor o perfil das publicações em periódicos, os 72 artigos publicados em periódicos foram categorizados segundo intervalos de fator de impacto (FI) — métrica que mensura a influência de um periódico com base na frequência de citações recebidas. A Figura 9 apresenta o gráfico da distribuição dos artigos selecionados na RSL, agrupados de acordo com o fator de impacto dos periódicos em que foram publicados.

Figura 9

Gráfico da distribuição das publicações selecionadas na Revisão Sistemática da Literatura do Estudo 1, categorizadas por intervalos de Fator de Impacto



Nota: Dados da pesquisa (2025).

A distribuição dos artigos revela que aproximadamente 60% das publicações concentram-se em periódicos com fator de impacto entre 1,01 e 5,0. Embora esse indicador não reflita com precisão a qualidade de artigos ou de pesquisadores individuais (Charan, 2014), ele serve como um parâmetro útil de visibilidade e influência moderadas na área. Os estudos selecionados evidenciam o panorama em constante evolução da inovação e da tecnologia, bem como os avanços na recuperação de informações impulsionada por IA, recursos que permitem a análise de grandes volumes de dados semiestruturados.

3.4.2 Análise das Informações Técnicas das Publicações Selecionadas na Revisão Sistemática da Literatura

A mineração de texto, definida como o processo de "descobrir informações em grandes coleções de texto e identificar automaticamente padrões e relacionamentos em dados textuais" (Feldman et al. 2007, p.1), é aplicada a documentos de patentes não apenas para extrair informações técnicas relacionadas ao conteúdo inventivo, mas também para identificar elementos que apoiam a gestão e a formulação de estratégias, com impacto direto na inovação, no desenvolvimento de negócios e na pesquisa.

A Tabela 8 sumariza as principais motivações que impulsionam os estudos sobre mineração de patentes, conforme identificadas nos artigos da RSL. Essas motivações estão categorizadas em quatro dimensões que se inter-relacionam e, em diversos casos, se

complementam.

Tabela 8

Síntese das principais motivações para a análise textual dos documentos de patente, categorizadas em dimensões

Dimensão	Descritivo	Principais artigos
Cognitiva	Coleta sistemática de inteligência técnica para apoiar atividades de pesquisa e desenvolvimento.	Berdyugina e Cavallucci (2021b, 2022a); Chen et al. (2020); Zheng et al. (2024); Jang e Yoon (2021)
	Identificação de informações técnicas generalizáveis, aplicáveis a diferentes contextos e setores.	
Tecnológica	Identificação de tecnologias novas, emergentes ou em declínio.	Joshi et al. (2022); Kim et al. (2019); Ryu e Lee (2024); Tian et al. (2022); Li et al. (2023)
	Detecção de lacunas tecnológicas relevantes para inovação e pesquisa.	
Mercado	Identificação de concorrentes.	Chuprat et al. (2024); Kim et al. (2019); Miao et al. (2022); Tian et al. (2022)
	Mapeamento de potenciais parceiros estratégicos.	
	Detecção de riscos de infração a direitos de propriedade intelectual.	
	Exploração de oportunidades estratégicas para inovação.	
	Análise de vantagens competitivas.	
	Identificação de patentes passíveis de compra ou licenciamento para acesso a novos mercados.	
Estratégica e Decisória	Orientação de investimentos.	Joshi et al. (2022); Liu et al. (2023); Park e Jun (2024); Puccetti et al. (2023); Kim et al. (2019)
	Definição de estratégias de propriedade intelectual.	
	Desenvolvimento de novos produtos e modelos de negócios.	
	Fornecimento de subsídios para tomada de decisão.	
	Apoio à formulação de políticas e estratégias eficazes de P&D.	

Nota: Elaborado pela Autora (2025).

As motivações identificadas se complementam ao evidenciar diferentes dimensões da mineração de patentes e, em conjunto, reforçam seu papel como instrumento estratégico para a inovação e para a gestão. A **dimensão cognitiva**, onde são extraídas e analisadas as informações inventivas, fornece a base de conhecimento técnico necessária para compreender o conteúdo das invenções, enquanto a **dimensão tecnológica** amplia essa dimensão ao situar as invenções em um panorama evolutivo, permitindo identificar tendências e lacunas. A **dimensão de mercado**, por sua vez, conecta a dimensão tecnológica ao contexto competitivo e econômico,

oferecendo subsídios para a identificação de concorrentes, parceiros e riscos associados. Finalmente, a **dimensão estratégica e decisória** integra os achados das dimensões anteriores e os transforma em diretrizes práticas, orientando investimentos, estratégias de propriedade intelectual e iniciativas de P&D. Dessa forma, as quatro dimensões não apenas se articulam de forma complementar, mas também demonstram como a mineração de patentes pode atuar como elo entre a geração de conhecimento técnico e sua aplicação em decisões estratégicas.

Com base nesse panorama, o estudo identificou sete estratégias de mineração de texto voltadas à extração de informações e conhecimentos significativos em grandes volumes de dados. Essas estratégias foram agrupadas em categorias conforme seu foco na análise em nível de palavra e/ou frase, bem como na consideração de contextos sintáticos e semânticos. A Tabela 9 sumariza as estratégias de mineração de texto, acompanhadas de breves descrições conceituais e de artigos representativos.

Tabela 9

Síntese das estratégias de mineração de texto e artigos relacionados

Categoria de análise	Estratégia de mineração textual	Breve descrição	Publicações
Análise conduzida no nível da palavra e/ou frase	Extração de palavras-chave e/ou frases	Essa estratégia concentra-se na identificação de palavras-chave técnicas e suas relações, embora possa ignorar conexões semânticas que transmitem desempenho e funcionalidade técnicos. Alguns estudos visam extrair frases técnicas ricas em informações, essenciais para a compreensão de textos de patentes.	Dessi et al. (2023); Hu et al. (2018); Hwang et al. (2022); Liu et al. (2023); Miao et al. (2022); Park e Jun (2024); Rossi et al. (2019); Russo et al. (2018); Sarica et al. (2019); Shin et al. (2023); Zhou et al. (2024); Liu et al. (2020).
	Análise de coocorrência	Esta técnica analisa as relações entre palavras para obter uma compreensão mais profunda do contexto, indo além da análise básica de palavras-chave. Ela se baseia em padrões de coocorrência para extrair relações de texto não estruturado, partindo do pressuposto de que entidades que aparecem juntas com frequência provavelmente estão relacionadas	Kim et al. (2020); Liu et al. (2023); Wang et al. (2024); Yue et al. (2023).

Categoria de análise	Estratégia de mineração textual	Breve descrição	Publicações
	Modelo de Espaço Vetorial	Essa abordagem representa cada patente como um vetor de palavras codificadas, transformando documentos de patentes em dados estruturados, atribuindo pesos aos termos com base em sua frequência.	Liu e Zhang et al. (2020); Wang e Liu (2022); Yue et al. (2023).
Análise sintática e semântica	Mineração Baseada em Conhecimento de Domínio	Essa estratégia envolve a construção de ontologias específicas de domínio para extrair informações tecnológicas de patentes de forma abrangente, embora isso frequentemente exija um esforço manual substancial. Essas ontologias definem atributos específicos de domínio e capturam interações entre subdomínios de conhecimento, fortalecendo assim a estrutura técnica para recuperação de patentes (Trappey et al., 2023).	Lin et al. (2022); Taduri et al. (2019); Trappey et al. (2018); Trappey, Trappey e Chang (2020); Trappey, Trappey, Wu, et al. (2020); Trappey et al. (2023).
	Mineração de Informação Semântica	Este método utiliza representações semânticas de textos de patentes para extrair informações usando regras de associação. O framework SAO facilita a mineração semântica, identificando oportunidades tecnológicas, com foco em problemas e funções técnicas. Como alternativa ao SAO, foi proposta a abordagem Função-Objeto-Propriedade, que extrai conhecimento semântico funcional (Teng et al., 2024).	Cui e Qian (2022); Jang et al. (2022); Jang e Yoon (2021); Korobkin, Fomenkov e Golovanchikov (2018); Korobkin, Fomenkov e Kolesnikov (2018); Li, Wang et al. (2023); Li et al. (2023); Lin et al. (2022); Sarica et al. (2020); Teng et al. (2024); Wang et al. (2024); Zhang et al. (2020).
	Modelagem de Tópicos	Esta abordagem estatística analisa as estruturas semânticas e sintáticas de textos para identificar padrões de tópicos latentes	Jiang et al. (2025); Krasnov et al. (2022); Ma et al. (2021); Nkolongo et al. (2024);

Categoria de análise	Estratégia de mineração textual	Breve descrição	Publicações
		em grandes coleções de textos. Métodos probabilísticos generativos, como Explicação de Correlação, Indexação Semântica Latente Probabilística e Modelos de Tópicos Dinâmicos são comumente utilizados para esse fim.	Tian et al. (2022, 2024); Trappey, Lin et al. (2024); Trappey et al. (2023); Wei et al. (2023).
	Extração de Propriedades, Funções, Entidades e Efeitos	Essa abordagem extrai termos técnicos multidimensionais, representando função, comportamento e estrutura, de textos de patentes por meio de um processo que reflete o processo de design do produto.	Ayaou et al. (2025; Berdyugina e Cavallucci (2021, 2022a, 2023); Chan et al. (2021); Deng, Chen, et al. (2018); Giordano et al. (2023); Hou et al. (2024); Kang et al. (2018); Kim, Joung, et al. (2018); Korobkin, Fomenkov, e Golovanchikov (2018); Korobkin, Fomenkov, e Kravets (2018); Korobkin et al. (2019); Kronemeyer et al. (2022); Li et al. (2024, 2025); Trapp e Warschat (2025); Trappey et al. (2024); Wang et al. (2024)

Nota: Elaborado pela Autora (2025).

Embora a análise semântica e a extração de palavras-chave sejam processos distintos, ambos são inter-relacionados e frequentemente complementares na mineração de texto. A extração de palavras-chave busca identificar os termos mais representativos do conteúdo central de um documento, enquanto a análise semântica procura compreender o significado desses termos e suas relações dentro do contexto textual. No caso da mineração de patentes, estratégias baseadas apenas na extração de palavras-chave podem negligenciar conexões semânticas, resultando em informações fragmentadas e na perda de significado contextual (Joshi et al.,

2022; Kim & Joung et al., 2018).

As técnicas de extração de palavras-chave constituem, portanto, um componente essencial da mineração de texto e, em muitos estudos, são utilizadas como etapa inicial de metodologias mais amplas que incorporam análises adicionais (Trappey et al., 2024; Zhao, 2024). Entre essas abordagens destacam-se: o uso da estrutura SAO para mapear relações funcionais (Kim et al., 2020); modelos semânticos como o BERT (*Bidirectional Encoder Representations from Transformers*), que convertem sentenças em representações vetoriais (Trappey et al., 2023; Wei et al., 2023); a construção de redes de conhecimento prévio para aprimorar a extração de termos (Huang & Xie, 2022); além da extração e reconhecimento de entidades nomeadas (Puccetti et al., 2023).

A integração entre mineração semântica e extração de palavras-chave aumenta significativamente a precisão e a relevância dos termos extraídos, ao incorporar informações contextuais e explorar relações semânticas. Evidências empíricas demonstram que a análise semântica favorece a recuperação de informações relevantes (Lin et al., 2022), melhora a precisão dos resultados (Kaliteevskii et al., 2021), auxilia na detecção de tecnologias emergentes e na previsão de tendências, além de mapear relações entre conceitos tecnológicos (Jang & Yoon, 2021; Kim et al., 2020). Também possibilita acompanhar a evolução tecnológica ao longo do tempo em textos de patentes (Lin et al., 2022; Sarica et al., 2020), oferecendo uma compreensão mais ampla e aprofundada do conteúdo (Kaliteevskii et al., 2021; Wei et al., 2023).

Mais recentemente, a integração entre mineração semântica e técnicas de inteligência artificial tem viabilizado análises gramaticais e sintáticas mais refinadas, potencializando a interpretação de conteúdos tecnológicos. Observa-se um crescimento expressivo do uso de PLN para extrair informações significativas de documentos de patentes, enquanto os métodos de ML e DL ampliam sua aplicação. Esse avanço é impulsionado pela necessidade de lidar com grandes volumes de dados de forma eficiente, permitindo análises escaláveis em múltiplos domínios tecnológicos (Ryu & Lee, 2024). Para ilustrar a evolução da área de pesquisa, a Tabela 10 apresenta a síntese dos avanços metodológicos mais significativos e os estudos que se destacam nesse campo, traçando um panorama temporal do desenvolvimento da área no período de 2018 a 2025.

Tabela 10

Síntese dos avanços metodológicos de mineração de patentes (2018–2025) e estudos representativos

Ano	Estudos representativos e principais avanços
2018	Extração e análise supervisionadas de palavras-chave e grupos temáticos (Hu et al., 2018). Extração de Estruturas SAO para a identificação de funções técnicas Korobkin, Fomenkov & Golovanchikov, 2018). Utilização de Redes Neurais Convolucionais (CNNs) e embeddings[4] de palavras para análise de texto (Li et al., 2018).
2019	Métodos que combinam ML e PLN para comparar palavras ou textos entre patentes (Helmers et al., 2019). SAO aplicado usando PLN (Kim et al., 2019). Uso de redes neurais recorrentes profundas (Wu, 2019).
2020	Definidas associações semânticas tecnicamente significativas entre palavras com base na coocorrência usando o algoritmo TechNet (Sarica et al., 2020). Associação de estratégias semânticas para extrair frases de efeito usando um método de DL, onde representações semânticas são aprendidas por meio de transformações vetoriais de palavras e relações de analogia são estabelecidas por meio de agrupamento de tópicos (Zhang & Yu, 2020). Modelo de Tópicos Heterogêneos para capturar termos que refletem significados semânticos distintos da mesma palavra em diferentes contextos (Chen et al., 2020). Análise de relações função-efeito na representação de patentes (Liu et al., 2020). Aplicações de PLN utilizando analisadores sintáticos para extrair recursos tecnológicos (Russo, 2020).
2021	Modelo semântico que utiliza técnicas de aprendizado de máquina e PLN para extrair informações sobre tendências evolutivas no desenvolvimento de conceitos (Kaliteevskii et al., 2021). Método de DL que utiliza BERT para extrair contradições dentro da estrutura TRIZ (Guarino et al., 2021). Utilização da rede neural BERT para extração de palavras-chave e transformação de palavras em representações vetoriai (Liu et al., 2021). Construção de um dicionário de terminologia de patentes com base em um modelo estatístico derivado de resumos de patentes chinesas (Wang et al., 2021).
2022	Extração de estruturas SAO que descrevem problemas e soluções tecnológicas, juntamente com a geração de vetores de documentos descritivos que consideram o contexto funcional dos termos tecnológico (Yoon et al., 2022). Utilização de Redes Adversariais Generativas (GANs) para obter resumos de patentes que aprendem informações técnicas considerando as características da patente (Kim & Yoon, 2022). Aplicação de um modelo DL para prever mudanças de termos ao longo do tempo (Hwang et al., 2022).
2023	Metodologia que integra TRIZ com PLN para formalizar e extrair princípios inventivos de documentos de patentes (Berdyugina & Cavallucci, 2023). Utiliza técnicas de aprendizado de máquina para classificar dados de texto não estruturados, capturando informações contextuais por meio de *embeddings* de texto

[4] *Embeddings* são vetores (uma matriz de números que define um ponto em um espaço dimensional) criados por ML com a finalidade de capturar dados significativos sobre cada objeto.

Ano	Estudos representativos e principais avanços
	(Miric et al., 2023). Combina DL com uma rede lexical construída pela análise das relações entre palavras e suas categorias associadas (Li, Yu et al., 2023).
2024	Construção de grafos de conhecimento para resumir e representar informações de pedidos de patente, particularmente por meio do uso de modelos de reconhecimento de entidade nomeada (NER - *Named Entity Recognition*) pré-treinados (Jeon et al., 2024). Construção de grafos de conhecimento ontológicos integrando modelos de PLN com agrupamento *K-means*, KeyBERT para extração de palavras-chave e CoreX para modelagem de tópicos (Trappey, Lin, et al., 2024). Aplicação de DL para capturar relacionamentos entre entidades (Zhang et al., 2024). Construção de redes semânticas usando LLMs (Giordano et al., 2024).
2025	Utilização do modelo *Google BERT for Patents*, pré-treinado em mais de 100 milhões de documentos de patentes, para capturar similaridades contextuais entre textos de patentes (Ali et al., 2025). Uso de LLMs para resumir automaticamente contradições técnicas a partir de dados textuais de patentes, identificar parâmetros técnicos relevantes e mapeá-los aos parâmetros de engenharia padronizados definidos pela metodologia TRIZ (Trapp & Warschat, 2025).

Nota: Elaborado pela Autora (2025).

Entre 2018 e 2019, a pesquisa em mineração de textos de patentes avançou fortemente na captura de relações semânticas. Em 2018, destacaram-se as primeiras aplicações da TRIZ como estrutura metodológica para análise de patentes (Korobkin, Fomenkov & Golovanchikov, 2018), além do uso inicial de técnicas de DL nesse campo (Lin et al., 2018). No ano seguinte, esses avanços foram ampliados com a proposta de métodos para construção automatizada de matrizes de funções técnicas baseadas em efeitos físicos (Korobkin et al., 2019), bem como pela aplicação integrada de ML e PLN para agrupamento semântico de termos (Kim et al., 2019) e comparação automática de textos completos com o objetivo de detectar invenções similares (Helmers et al., 2019). Técnicas de SAO (Kim et al., 2019) e redes neurais recorrentes (Wu, 2019) também começaram a ser exploradas para identificar padrões funcionais e características latentes em títulos e resumos de patentes.

Em 2020, a ênfase deslocou-se para o tratamento de relações lexicais e semânticas mais amplas. Surgiram o algoritmo Technet, para recuperação de termos interconectados (Sarica et al., 2020), e o Modelo de Tópicos Heterogêneos (Chen et al., 2020), que combina *embeddings* de palavras e representações de patentes para capturar expressões semanticamente equivalentes. Nesse mesmo período, Russo (2020) propôs uma abordagem baseada em PLN e análise de dependências para automatizar a extração de funções, aplicações e requisitos tecnológicos.

Entre 2021 e 2022, os esforços concentraram-se na extração automática de informações

estruturadas. Foram propostas abordagens para criação automática de corpora de treinamento para NER **(**Puccetti et al., 2021), geração de dicionários terminológicos para patentes chinesas **(**Wang et al., 2021), e detecção de contradições técnicas via modelos como o BERT (Guarino et al., 2021). Também se destacaram o TechWordNet (Jang e Yoon, 2021) e a combinação de LDA, DTM e LSTM para mapear evolução terminológica e prever tendências (Hwang et al., 2022). Adicionalmente, estudos passaram a enfatizar a unificação de expressões funcionais e a construção de dicionários de sinônimos e hipônimos tecnológicos (Shi et al., 2022).

Os anos de 2023 e 2024 marcaram uma forte consolidação de abordagens não supervisionadas e fracamente supervisionadas, motivadas pela escassez de dados anotados. O método TechPat (Liu et al., 2023) e o modelo DeepKea (Dessi et al., 2023) exemplificam esse movimento, explorando desde marcação gramatical até redes neurais profundas com *embeddings* pré-treinados. No campo de NER, estudos compararam métodos baseados em dicionários, regras e aprendizado estatístico, destacando o melhor desempenho dos modelos baseados em ML (Puccetti et al., 2023). Paralelamente, ontologias derivadas da TRIZ, como o IDM (Berdyugina & Cavallucci, 2023), reforçaram a extração de conhecimento estruturado. Já em 2024, consolidou-se a aplicação de grafos de conhecimento apoiados em técnicas de extração como o KeyBERT (Trappey, Lin, et al., 2024), NER+NEN (*Named Entity Normalization*) (Jeon et al., 2024) e modelos híbridos que combinam aprendizado estatístico e DL (Zhang et al., 2024; Zhou et al., 2024)

Em 2025, a literatura destacou o uso de modelagem dinâmica de tópicos para detectar sinais emergentes, ainda que desafiados por ruído e terminologia altamente especializada (Jiang et al., 2025). Nesse cenário, os LLMs surgem como alternativa promissora, capazes de capturar padrões contextuais sofisticados e de apoiar tarefas como sumarização de contradições técnicas e mapeamento de parâmetros TRIZ (Trapp & Warschat, 2025). Complementarmente, modelos especializados como o Google BERT *for Patents*, treinado em cerca de 8.000 termos técnicos, foram empregados para lidar com o vocabulário jurídico e técnico específico das patentes (Ali et al., 2025).

De forma geral, a linha do tempo revela uma trajetória em que a mineração textual de patentes evolui de abordagens léxico-estatísticas e modelos probabilísticos (como LDA) para métodos cada vez mais semânticos, contextuais e baseados em DL/LLMs. Essa progressão demonstra um movimento claro rumo ao refinamento metodológico e ao aumento da capacidade interpretativa, com forte integração de ontologias, PLN e aprendizado profundo como pilares da próxima geração de sistemas de inteligência técnica.

3.4.3 Mapeamento das Contribuições e Propostas de Pesquisas a partir dos Artigos

Com o objetivo de identificar tópicos inexplorados ou pouco pesquisados, bem como descobrir soluções alternativas para orientar pesquisas futuras, foram analisados os *insights* provenientes das contribuições e das propostas de direções futuras. Os resultados foram organizados em tópicos-chave, incluindo: campos não estruturados dos documentos de patentes explorados, a fase de pré-processamento, o processo de mineração de texto e o uso de métricas de desempenho.

Campos de Patentes Explorados

Não há consenso sobre qual campo não estruturado dentro de um documento de patente é mais apropriado para análise. Embora a seção de descrição ofereça mais detalhes técnicos sobre a tecnologia (Wang & Liu, 2022), ela costuma ser extensa e contém ruído significativo (Sarica et al., 2020; Son et al., 2022). Alguns pesquisadores consideram a seção de reivindicações a principal fonte de informações inventivas. No entanto, a mistura de linguagem jurídica e técnica pode introduzir termos ruidosos (Berdyugina & Cavallucci, 2020a; Chen et al., 2020; Geng, 2021). Para abordar isso, alguns sugerem focar apenas na reivindicação independente, que define os limites da invenção (Joshi et al., 2022). Em relação ao resumo, Wang e Liu (2024) alertam que seu formato condensado pode reduzir a precisão ou introduzir viés, pois informações importantes podem ser omitidas. Eles propõem que a combinação do resumo com as reivindicações e a descrição técnica produz uma análise mais abrangente. Por outro lado, Son et al. (2022) argumentam que confiar apenas em resumos e reivindicações pode não fornecer detalhes suficientes sobre o cerne técnico da invenção. Eles observam que essas seções frequentemente apresentam apenas uma parte do conteúdo técnico da invenção e tendem a enquadrá-la em termos gerais e jurídicos.

Pré-processamento do texto bruto

No pré-processamento de texto, termos não informativos, comumente chamados de *stopwords* (ou palavras de interrupção), são removidos, juntamente com outras palavras e sinais de pontuação irrelevantes, para reduzir o ruído nos dados (Berdyugina & Cavallucci, 2023; Joshi et al., 2022; Maskittou et al., 2022). Bibliotecas genéricas de *stopwords* são normalmente empregadas em tarefas de PLN em diversos domínios, como as disponíveis no *NLTK* (Sarica et al., 2019). No entanto, essas bibliotecas frequentemente exigem personalização quando aplicadas a textos de patentes. Termos recorrentes específicos desse domínio, como “invenção”, “reivindicação” e “figura”, também devem ser filtrados, pois sua presença pode comprometer a qualidade da extração de palavras-chave e de outros resultados analíticos (Shin et al., 2023;

Xiao et al., 2018).

As etapas padrão de pré-processamento também incluem a tokenização e a conversão para letras minúsculas, a fim de garantir a uniformidade do conjunto de dados (Chen et al., 2020; Vereschak & Korobkin, 2019). A qualidade da tokenização é particularmente crítica, pois influencia significativamente o desempenho de modelos de linguagem em contextos específicos de domínio (Althammer et al., 2021). Além disso, formas flexionadas das palavras são tratadas por meio da lematização, que as reduz às suas formas básicas (Berdyugina & Cavallucci, 2023). Ferramentas como NLTK e *SpaCy* são amplamente utilizadas em diversas tarefas de pré-processamento, incluindo tokenização e marcação de classes gramaticais (Tian et al., 2022; Wang and Liu, 2024).

Estratégia de Mineração de Texto

Para lidar com os desafios lexicais e semânticos inerentes aos documentos de patentes, como a variabilidade linguística, o uso de sinônimos e antônimos, e a terminologia específica de domínio, diversas estratégias vêm sendo identificadas. No contexto da mineração semântica, é essencial representar a semântica contextual (Wei et al., 2023), bem como identificar a estrutura sintática dos termos extraídos (Wang & Liu, 2024) ou as características estruturais das sentenças (Zhou et al., 2024), a fim de capturar de forma eficaz as informações contextuais e semânticas (Dessi et al., 2023; Lv et al., 2019; Wang et al., 2024; Zhang et al., 2022; Zheng et al., 2024).

Para a implementação de algoritmos de extração, é recomendado o uso de técnicas de PLN e de LLMs. Modelos como Word2Vec, BERT, GPT-3 (Giordano et al. 2023, 2024) e LLMs mais especializados e refinados (Blume et al., 2024) são empregados para capturar relacionamentos implícitos que não estão explicitamente declarados em um único documento. O Word2Vec, desenvolvido pelo *Google*, representa palavras como vetores de valores reais, codificando eficientemente informações semânticas e mitigando problemas relacionados ao conhecimento lexical limitado (Chen et al., 2020; Chiarello et al., 2018; Jing et al., 2023; Lee et al., 2022; Lu et al., 2019; Ma et al., 2018; Yoon et al., 2022).

Para tarefas de PLN que envolvem compreensão contextual, o modelo RoBERTa, uma versão otimizada do BERT, é recomendado devido ao seu desempenho superior na captura de estruturas linguísticas complexas e dependências contextuais (Liang et al., 2024; Nkolongo et al., 2024; Trapp et al., 2023). Em tarefas de mineração em conjuntos de dados de domínio único, o Patent-BERT demonstra maior robustez semântica e melhor adaptação ao domínio de destino, em comparação ao modelo *BERT-for-Patents* do Google (Wang et al., 2024).

No que diz respeito ao NER, abordagens tradicionais podem ser insuficientes, pois

geralmente exigem amplo conhecimento do domínio e grande esforço de engenharia para identificar categorias predefinidas de objetos no texto, tornando o processo demorado (Zhang & Yu, 2020), dispendioso e suscetível a erros (Hu et al., 2018). Dada a complexidade estrutural e a linguagem densa dos documentos de patentes, são recomendados métodos alternativos de NER (Puccetti et al., 2021; Saad, 2019)

Para enfrentar os desafios lexicais e semânticos, decorrentes da variabilidade linguística, incluindo sinônimos, antônimos e jargões técnicos, foram propostas estratégias como o uso de representações linguísticas adaptadas ao domínio em métodos de pré-treinamento (Althammer et al., 2021), a criação de dicionários terminológicos específicos (Wang et al., 2021), a associação entre efeitos, parâmetros e propriedades (Zhai et al., 2020) e o desenvolvimento de ontologias (Phan et al., 2018; Russo et al., 2018). Tais abordagens visam revelar relações semânticas e, assim, aprimorar a adaptabilidade das técnicas de mineração de texto à linguagem especializada empregada em patentes.

Nesse contexto, ontologias semânticas extraídas da TRIZ, baseadas em efeitos físicos, princípios inventivos e contradições, têm sido propostas para a extração automática de informações inventivas em textos de patentes (Berdyugina & Cavallucci, 2020a, 2022b, 2023; Kang et al., 2018; Korobkin, Fomenkov, & Kravets, 2018; Trapp & Warschat, 2025; Wang et al., 2024; Yun et al., 2022). Igualmente importante, o tratamento de nuances semânticas e sintáticas em múltiplos idiomas continua sendo um desafio central na mineração de textos de patentes, sobretudo devido à natureza inerentemente multilíngue dos bancos de dados de patentes (Kim et al., 2020; Liwei, 2022; Wang et al., 2021).

Métricas de desempenho

Métricas de desempenho aprimoradas em metodologias de mineração de texto foram observadas em domínios mais específicos, nos quais os documentos de patentes apresentam maior consistência e similaridade (Trappey, Trappey, Wu, et al., 2020). Em contraste, em diversos domínios técnicos, variações textuais introduzem distinções que podem impactar negativamente essas métricas (Chen et al., 2022; Fink et al., 2021). Para mitigar esse problema, estudos recomendam a incorporação de vocabulário específico de patentes do domínio em questão, com o objetivo de aumentar a eficácia da extração de informações (Geng, 2021).

3.5 Discussão

Os resultados desta Revisão Sistemática da Literatura evidenciam que a mineração

textual de documentos de patentes evoluiu significativamente nos últimos anos, acompanhando a crescente complexidade dos sistemas de inovação e a intensificação da produção tecnológica global. Embora as bases de patentes concentrem grandes volumes de informação técnica, os estudos analisados convergem ao demonstrar que esse potencial informacional permanece subexplorado na ausência de métodos robustos de mineração, recuperação e interpretação do conteúdo textual.

Sob a perspectiva da KBV, os documentos de patentes configuram-se como importantes ativos de conhecimento codificado. No entanto, seu valor estratégico não reside apenas na posse da informação, mas na capacidade organizacional de transformá-la em conhecimento acionável. Nesse sentido, os métodos de mineração textual atuam como mecanismos intermediários que viabilizam a conversão de informação técnica dispersa em conhecimento estruturado, passível de integração aos processos de P&D, inovação e tomada de decisão estratégica.

A análise das tendências de publicação revela um deslocamento progressivo de conferências para periódicos científicos, especialmente a partir de 2022, indicando um estágio de maior maturidade teórica e metodológica do campo. Esse movimento sugere o fortalecimento da capacidade absortiva da comunidade científica, refletindo avanços nos processos de assimilação e transformação do conhecimento previamente adquirido. Adicionalmente, a predominância de países asiáticos, em especial China e Coreia do Sul, pode ser interpretada como reflexo de ecossistemas nacionais com elevada capacidade de absorver, recombinar e explorar conhecimento tecnológico oriundo de patentes, em consonância com o aumento expressivo dos pedidos de patente nessas economias.

No plano metodológico, os resultados indicam uma transição clara de abordagens baseadas em regras manuais para técnicas fundamentadas em aprendizado de máquina, aprendizado profundo e processamento de linguagem natural. Enquanto métodos rule-based exigem elevado conhecimento de domínio e apresentam baixa escalabilidade, técnicas baseadas em DL e LLMs demonstram maior capacidade de adaptação à linguagem técnica, jurídica e multilíngue das patentes. Esse avanço amplia significativamente a capacidade de aquisição e assimilação do conhecimento externo, elementos centrais da capacidade absortiva.

A literatura também destaca a importância da análise semântica e da modelagem ontológica como estratégias para mitigar limitações associadas à variabilidade linguística, à polissemia e ao uso intensivo de sinônimos em textos de patentes. Nesse contexto, ontologias específicas de domínio emergem como instrumentos fundamentais para estruturar o conhecimento tecnológico. A TRIZ apresenta-se como um arcabouço particularmente relevante, ao fornecer conceitos formalizados de problemas, soluções, parâmetros e

contradições técnicas, compatíveis com regras sintáticas e semânticas aplicáveis à linguística computacional. Sua integração com técnicas de mineração textual fortalece tanto a interpretação do conteúdo inventivo quanto sua generalização para diferentes contextos tecnológicos.

3.6 Considerações Finais

Este estudo teve como objetivo consolidar e analisar criticamente a literatura recente sobre mineração textual de documentos de patentes, com ênfase nos avanços metodológicos, nas motivações de pesquisa e nas estratégias empregadas para extração de conhecimento tecnológico. Como principal contribuição acadêmica, a RSL oferece uma visão estruturada da evolução do campo entre 2018 e 2025, identificando tendências consolidadas, lacunas persistentes e direções promissoras para pesquisas futuras.

Os resultados reforçam que a inteligência técnica extraída de patentes desempenha papel estratégico não apenas na proteção da propriedade intelectual, mas também no apoio à inovação, à vigilância tecnológica e à formulação de estratégias de P&D. Nesse sentido, a mineração textual constitui um elemento-chave para operacionalizar os pressupostos da KBV, ao permitir que informações técnicas sejam internalizadas, articuladas ao conhecimento prévio e transformadas em vantagem competitiva. Do mesmo modo, os achados evidenciam que o avanço de técnicas semânticas, ontológicas e baseadas em DL contribui diretamente para o fortalecimento da capacidade absortiva, ao reduzir barreiras cognitivas e técnicas à exploração do conhecimento externo.

Entre as limitações do estudo, destaca-se a possibilidade de viés decorrente da estratégia de busca e dos critérios de seleção adotados na RSL, que podem ter levado à exclusão de estudos relevantes. Ainda assim, os resultados permitem delinear recomendações consistentes para pesquisas futuras, incluindo: (i) o aprofundamento da modelagem semântica e sintática em diferentes domínios tecnológicos; (ii) o desenvolvimento de ontologias específicas, alinhadas lexical e semanticamente à linguagem das patentes; (iii) a ampliação de abordagens para mineração multilíngue, especialmente em idiomas ainda pouco explorados, como o português e o espanhol; e (iv) a integração de bases terminológicas com repositórios científicos dinâmicos, visando à atualização contínua do vocabulário tecnológico.

Sob a perspectiva prática e gerencial, este estudo oferece subsídios relevantes para pesquisadores, gestores de inovação e formuladores de políticas públicas, ao demonstrar como a mineração textual de patentes pode apoiar decisões estratégicas, antecipar tendências

tecnológicas e identificar oportunidades emergentes. Em conjunto, as contribuições apresentadas reforçam o papel da mineração de textos de patentes como um campo em consolidação, com elevado potencial para o desenvolvimento de métodos mais eficazes, semanticamente contextualizados e aplicáveis a múltiplos domínios tecnológicos.

4 ESTUDO 2: INTEGRAÇÃO ENTRE TRIZ E MINERAÇÃO DE TEXTOS DE PATENTES: AVANÇOS, DESAFIOS E TENDÊNCIAS NA EXTRAÇÃO DE INTELIGÊNCIA TÉCNICA

Resumo

As patentes constituem uma das principais fontes de inteligência técnica, por concentrarem conhecimento científico e tecnológico codificado. Sob a perspectiva da Visão Baseada no Conhecimento (KBV), esses documentos representam ativos estratégicos cujo valor depende da capacidade organizacional de reconhecer, assimilar e aplicar o conhecimento neles contido. Contudo, o elevado volume de documentos e a complexidade linguística dos textos de patentes impõem desafios relevantes à recuperação eficaz de informações. Nesse contexto, a Teoria da Resolução Inventiva de Problemas (TRIZ) oferece um conjunto sistemático de ferramentas conceituais para a extração de inteligência técnica a partir de textos de patentes. A mineração textual e a TRIZ são abordagens complementares, pois compartilham o objetivo de reduzir barreiras cognitivas à resolução de problemas e de apoiar a identificação de soluções inventivas em múltiplos domínios tecnológicos. Ao estruturar problemas, soluções, parâmetros e contradições, a TRIZ contribui para transformar informação técnica em conhecimento organizacional estruturado. Este estudo investiga de que forma a TRIZ tem influenciado a evolução metodológica dos estudos sobre mineração de textos de patentes. Para isso, analisa-se a tendência temporal das publicações que exploram essa integração, identificando avanços e desafios com vistas à construção de uma agenda de pesquisa voltada ao desenvolvimento de um método que associe TRIZ e mineração textual para a análise de patentes. A pesquisa baseia-se em uma revisão sistemática da literatura composta por 75 artigos de periódicos e conferências, publicados em inglês até setembro de 2025. Os resultados indicam que essa integração constitui uma área promissora, marcada pelo uso crescente de aprendizado de máquina e aprendizado profundo, fortalecendo a capacidade absortiva e apoiando os processos de decisão e inovação.

Palavras-chave: Mineração textual. Patente. TRIZ. Inteligência técnica. Revisão sistemática da literatura.

Abstract

Patents constitute one of the main sources of technical intelligence, as they concentrate codified scientific and technological knowledge. From the perspective of the Knowledge-Based View (KBV), these documents represent strategic assets whose value depends on an organization's ability to recognize, assimilate, and apply the knowledge they contain. However, the large volume of documents and the linguistic complexity of patent texts pose significant challenges to effective information retrieval. In this context, the Theory of Inventive Problem Solving (TRIZ) provides a systematic set of conceptual tools for extracting technical intelligence from patent texts. Text mining and TRIZ are complementary approaches, as they share the objective of reducing cognitive barriers to problem solving and supporting the identification of inventive solutions across multiple technological domains. By structuring problems, solutions, parameters, and contradictions, TRIZ contributes to transforming technical information into structured organizational knowledge. This study investigates how TRIZ has influenced the methodological evolution of research on patent text mining. To this end, it analyzes the temporal trends of publications that explore this integration, identifying advances and

challenges with the aim of building a research agenda focused on the development of a method that combines TRIZ and text mining for patent analysis. The study is based on a systematic literature review comprising 75 journal and conference papers published in English up to September 2025. The results indicate that this integration represents a promising research area, characterized by the increasing use of machine learning and deep learning approaches, strengthening absorptive capacity and supporting decision-making and innovation processes.

Keywords: *Text mining. Patent. TRIZ. Technological intelligence. Systematic literature review.*

4.1 Introdução

A aquisição de informações científicas e tecnológicas, que compõem a inteligência técnica (Behkami & Daim, 2012), é um processo essencial para o avanço da Ciência, Tecnologia e Inovação (CT&I). Sob a perspectiva da Visão Baseada no Conhecimento (KBV), esse tipo de informação constitui um ativo estratégico, cujo valor não reside apenas em sua disponibilidade, mas na capacidade das organizações de reconhecê-lo, assimilá-lo e aplicá-lo de forma efetiva (Cohen & Levinthal, 1990). No âmbito da gestão de negócios, a democratização do acesso a conhecimentos de natureza técnica ou tecnológica pode impulsionar o desenvolvimento tecnológico (Belenzon, 2012; Han et al., 2006), ampliar o potencial inovador das economias modernas (Viglioni et al., 2023) e fortalecer a pesquisa científica e tecnológica.

Nesse contexto, a capacidade absortiva, definida por Zahra e George (2002) como a habilidade organizacional de adquirir, assimilar, transformar e explorar conhecimento externo, desempenha papel central na conversão da informação técnica em conhecimento organizacional e em resultados inovadores. Além disso, o intercâmbio entre diferentes domínios do conhecimento, por meio de processos de recombinação, transferência e transformação, constitui uma estratégia fundamental para a pesquisa e o desenvolvimento (Liu, Li, et al., 2020) ao ampliar a base de conhecimento disponível e potencializar a geração de soluções tecnológicas. Assim, a articulação entre inteligência técnica, KBV e capacidade absortiva reforça a importância de mecanismos sistemáticos de aquisição e integração do conhecimento para sustentar a inovação e a competitividade organizacional.

Como uma das principais fontes de inteligência técnica (Liwei, 2022; Xu et al., 2022), as bases patentárias revelam tecnologias recentes e avançadas, para uma variedade de domínios tecnológicos (Deng, Wang, et al., 2018; Krestel, Aras, et al., 2021). No entanto, essas informações inventivas são pouco exploradas no meio acadêmico (Mafu, 2023; Pimenta, 2017; Resende Ferreira et al., 2022), na indústria (Heisig et al., 2020; McTeague & Chatzimichali, 2022) e em *startups*, onde seria esperado o uso intensivo de bases de dados de patentes (Mazieri

et al., 2016). Dentre as possíveis causas podem ser citados o grande volume de dados, muito superior à capacidade humana de leitura (Chiarello et al., 2018) e a pouca familiaridade com a estrutura e a complexidade dos textos, com a presença de termos técnicos e jurídicos (Berdyugina & Cavallucci, 2021; Kim & Yoon, 2022; Puccetti et al., 2023).

Para extrair informações de grandes volumes de texto, facilitando o acesso à informação (Chiarello et al., 2018), várias abordagens de mineração textual têm sido usadas: a análise baseada em palavras-chave (Tseng et al., 2007; Yoon & Park, 2004); a abordagem SAO, formada pelas combinações gramaticais "substantivo-verbo-substantivo" que representa implicitamente uma solução (sujeito) e função (ação-objeto) de uma tecnologia (Korobkin, Fomenkov, & Golovanchikov, 2018); e a análise de propriedade-função (PF) (Dewulf, 2011), onde o núcleo da tecnologia é caracterizado pelos fatores propriedades (expresso como "adjetivos-substantivos") e funções (expresso como verbo-substantivo) (Kim, Choi, et al., 2018).

Entretanto, estudos na área têm evidenciado problemas de eficiência, especialmente quanto à definição dos termos de busca mais adequados para acessar determinados domínios de conhecimento. Assim, para a extração de terminologias específicas, que se aproximam da estrutura dos documentos de patente — nos quais são apresentados problemas e respectivas soluções —, diversos estudos têm se dedicado ao desenvolvimento de métodos baseados nas ferramentas da TRIZ (Berdyugina & Cavallucci, 2021, 2022b, 2023; Guarino et al., 2021; Kang et al., 2018; Kim et al., 2019; Kim & Yoon, 2022; Korobkin et al., 2019; Korobkin, Fomenkov, & Golovanchikov, 2018; Yue, Liu, Hou, et al., 2023).

Buscando contribuir para essa literatura, este estudo apresenta os resultados de uma revisão sistemática da literatura sobre a adoção das ferramentas e conceitos da TRIZ na mineração de textos de patentes. O objetivo é analisar a tendência temporal e metodológica das publicações que exploram a integração entre TRIZ e mineração de textos de patentes, identificando avanços e desafios da área, com vistas à formulação de uma agenda de pesquisa para o desenvolvimento de um método que associe TRIZ e mineração textual para a extração de inteligência técnica de patentes.

O estudo está estruturado em cinco seções principais: a primeira apresenta a revisão da literatura; a segunda descreve os procedimentos metodológicos adotados na revisão sistemática; a terceira expõe a análise dos resultados; a quarta reúne as discussões e a quinta apresenta as considerações finais.

4.2 Revisão da Literatura

A inteligência técnica em patentes constitui uma fonte estratégica de conhecimento sobre inovações e soluções tecnológicas, cuja análise pode ser potencializada por meio das abordagens de mineração textual, capazes de extrair automaticamente informações relevantes de grandes volumes de dados. Nesse cenário, a Teoria da Resolução Inventiva de Problemas (TRIZ) oferece uma estrutura conceitual sistemática que orienta a identificação de princípios e padrões inventivos, possibilitando uma compreensão mais profunda das relações tecnológicas e das soluções descritas nas patentes.

4.2.1 Inteligência Técnica em Patentes e Abordagens de Mineração Textual

Nos processos de pesquisa, desenvolvimento e inovação (P,D&I), a utilização das informações científicas e tecnológicas constitui a Inteligência Técnica (Behkami & Daim, 2012). A inteligência técnica estrutura os atributos tecnológicos em palavras para descrever uma tecnologia que se dedica à resolução de um problema técnico (Yoon & Park, 2004). Esses atributos compreendem a informação tecnológica, onde são descritos os recursos tecnológicos, sob a forma de produtos e processos (Jang et al., 2021), e a informação técnica, relacionada aos recursos intangíveis, notadamente conhecimentos, habilidades e ideias que são utilizados pelos gestores para as atividades organizacionais (Ali et al., 2020).

Os repositórios de patente são uma das principais fontes de inteligência técnica (Liwei, 2022; Pimenta, 2017; Xu et al., 2022), revelando tecnologias recentes e avançadas, para uma variedade de domínios tecnológicos (Deng, Wang, et al., 2018; Krestel, Chikkamath, et al., 2021). Este conhecimento técnico documentado em patentes tem uma forma específica de expressão (Sun et al., 2021), sendo previstas informações de metadados, sob a forma de dados estruturados, e dados não estruturados no formato textual e/ou imagens. Os dados estruturados são consistentes em semântica e formato, compreendendo as informações bibliográficas, tal como titular da patente, inventores e um classificador que categoriza as patentes em campos tecnológicos (Krestel et al., 2021; Sun et al., 2022). Para a análise dos campos estruturados, geralmente são utilizados estudos métricos, permitindo identificar tendências e padrões tecnológicos e orientar a tomada de decisões (Ki & Kim, 2017). Os campos não estruturados, que incluem o título, a descrição detalhada da invenção, as reivindicações, resumo e figuras, não estão organizados em um modelo predefinido (Zanella et al., 2023).

Para extrair a descrição das funções técnicas do texto de patente, as abordagens baseadas

em léxicos têm sido insuficientes para refletir conceitos-chave tecnológicos específicos (Choi et al., 2013; Park, Ree, et al., 2013). Isto porque os campos não estruturados de patentes frequentemente apresentam termos com características gerais da terminologia de Ciência e Tecnologia, cujos princípios básicos incluem termos chave, simples ou complexo, relacionados a um determinado domínio da tecnologia (Liwei, 2022). Estes termos se relacionam com atributos de solução e função, que expressam conceitos-chave de uma tecnologia (Kim, Choi, et al., 2018).

Nesse contexto, a partir dos anos 2000, diversos estudos de mineração de textos de patentes passaram a realizar a extração de termos relacionados a problemas técnicos e soluções, utilizando as ferramentas da TRIZ (Cascini & Russo, 2007; Liang et al., 2008; Soo et al., 2005). Esses conceitos fundamentais estão diretamente ligados ao processo inventivo, geralmente entendido como uma atividade de resolução de problemas que envolve a identificação e definição do problema, a geração e avaliação de soluções alternativas e a seleção da alternativa mais promissora (Giordano et al., 2023).

4.2.2 Fundamentos e Aspectos Conceituais da TRIZ

A TRIZ, acrônimo da expressão russa Teorija Rezhenija Izobretatelskih Zadach (Teoria da Resolução de Problemas Inventivos), foi desenvolvida por Genrich Altshuller na União Soviética entre 1946 e 1985 (Hmina et al., 2019; Ilevbare et al., 2013). Embora frequentemente denominada como uma teoria, a TRIZ pode ser mais adequadamente compreendida como uma metodologia sistemática voltada à resolução de problemas tecnológicos, baseada na análise empírica de soluções inventivas registradas em patente (Savransky, 2000; Yan et al., 2015).

A formulação da TRIZ resultou da análise de aproximadamente 400.000 documentos de patente, a partir dos quais Altshuller identificou regularidades, padrões recorrentes e princípios subjacentes aos processos de resolução de problemas técnico (Kim et al., 2009). Um dos achados centrais dessa análise foi a constatação de que apenas cerca de 0,3% das soluções patenteadas eram verdadeiramente disruptivas, isto é, fundamentadas em princípios físicos recém-descobertos. Os 99,7% restantes baseavam-se em princípios já conhecidos, distinguindo-se pela forma inovadora como esses princípios eram combinados, adaptados ou implementados (Souchkov, 2017).

Com base nesses padrões recorrentes, Altshuller sintetizou um conjunto de 40 Princípios Inventivos, que representam soluções genéricas aplicáveis a uma ampla variedade de problemas técnicos (Carvalho & Back, 2001; Cong & Tong, 2008). Esses princípios podem ser compreendidos como uma abstração das soluções encontradas nas patentes analisadas,

constituindo a base conceitual de aproximadamente 80% das invenções patenteadas (Lim, 2016; Lux, 2022). Paralelamente, os problemas técnicos foram organizados em 39 parâmetros de engenharia, que descrevem características técnicas frequentemente envolvidas em conflitos de projeto. A Tabela 11 apresenta os 39 parâmetros da TRIZ.

Tabela 11

Parâmetros de Engenharia TRIZ

1	Peso de um objeto em movimento	2	Potência
3	Peso de um objeto sem movimento	4	Perda de energia
5	Comprimento de um objeto em movimento	6	Perda de substância
7	Comprimento de um objeto sem movimento	8	Perda de informação
9	Área de um objeto em movimento	10	Perda de tempo
11	Área de um objeto sem movimento	12	Quantidade de substância
13	Volume de um objeto em movimento	14	Confiabilidade
15	Volume de um objeto sem movimento	16	Certeza de medição
17	Velocidade	18	Certeza de manufatura
19	Força	20	Fatores danosos atuando sobre o objeto
21	Tensão/Pressão	22	Efeitos colaterais danosos
23	Forma	24	Fabricabilidade
25	Estabilidade do objeto	26	Conveniência de uso
27	Resistência	28	Reparabilidade
29	Durabilidade de um objeto em movimento	30	Adaptabilidade
31	Durabilidade de um objeto sem movimento	32	Complexidade do objeto
33	Temperatura	34	Complexidade de controle
35	Brilho	36	Grau de automação
37	Energia gasta por um objeto em movimento	38	Produtividade
39	Energia gasta por um objeto sem movimento		

Nota. Obtido de Navas (2013).

Altshuller verificou que todas as inovações emergem de um número reduzido de princípios inventivos (Liang et al., 2008). O criador da TRIZ constatou que toda solução inovadora resulta da eliminação de uma contradição, e que apenas um conjunto limitado de

princípios inventivos era utilizado para resolver ou eliminar essas contradições, conforme evidenciado na análise de milhares de invenções, independentemente da área tecnológica das patentes (Souchkov, 2016, 2017). A Tabela 12 apresenta os 40 Princípios Inventivos da TRIZ.

Tabela 12

Princípios Inventivos TRIZ

1	Segmentação	21	Aceleração
2	Extração	22	Transformação do prejuízo em lucro
3	Qualidade localizada	23	Feedback
4	Assimetria	24	Mediação
5	Consolidação	25	Auto-serviço
6	Universalidade	26	Cópia
7	Aninhamento	27	Uso e descarte
8	Contrapeso	28	Substituição de meios mecânicos
9	Compensação prévia	29	Construção pneumática ou hidráulica
10	Ação prévia	30	Uso de filmes finos e membranas flexíveis
11	Amortecimento prévio	31	Uso de materiais porosos
12	Equipotencialidade	32	Mudança de cor
13	Inversão	33	Homogeneização
14	Recurvação	34	Descarte e regeneração
15	Dinamização	35	Mudança de estado físico ou químico
16	Ação parcial ou excessiva	36	Mudança de fase
17	Transição para nova dimensão	37	Expansão térmica
18	Vibração mecânica	38	Uso de oxidantes fortes
19	Ação periódica	39	Uso de atmosferas inertes
20	Continuidade da ação útil	40	Uso de materiais compostos

Nota. Obtido de Navas (2013).

Esses parâmetros e princípios foram integrados na Matriz de Contradições da TRIZ**,** uma de suas ferramentas centrais. Na perspectiva da TRIZ, uma contradição técnica ocorre quando a melhoria de uma característica do sistema implica a degradação de outra, configurando-se como um dos principais obstáculos à solução de problemas inventivos (Liang et al., 2008; Srinivasan & Kraslawski, 2006). A Matriz de Contradições foi concebida para orientar o processo criativo, relacionando pares de parâmetros conflitantes aos princípios inventivos mais adequados para sua superação (Yan et al., 2015).

Os conceitos centrais da TRIZ (problemas técnicos, soluções técnicas, contradições e princípios inventivos) estão intrinsecamente relacionados ao processo inventivo descrito nos documentos de patente. Esses documentos apresentam soluções para problemas técnicos utilizando uma terminologia científica e tecnológica relativamente padronizada, associada a domínios específicos do conhecimento (Liwei, 2022). Nesse sentido, as patentes constituem não apenas um repositório jurídico, mas também uma base estruturada de conhecimento técnico, cuja análise sistemática permite a identificação de padrões, soluções genéricas e princípios aplicáveis a diferentes contextos tecnológicos.

Ao buscar soluções inovadoras em diferentes campos da ciência e da engenharia, a TRIZ inspirou a criação das bases de efeitos, que reúnem conceitos extraídos do conhecimento científico e da engenharia, aplicados à resolução de problemas (Ilevbare et al., 2013). Esses efeitos físicos podem ser empregados na solução de problemas fora do domínio em que foram originalmente identificados. A resolução de problemas complexos também pode ser facilitada pela combinação de múltiplos efeitos (Navas, 2013). Assim como outras ferramentas da TRIZ, as bases de efeitos físicos sugerem soluções com base em diversas aplicações possíveis de cada efeito, sendo escolhidas aquelas que se alinham ao contexto de uma aplicação específica (Gao & Zhu, 2015).

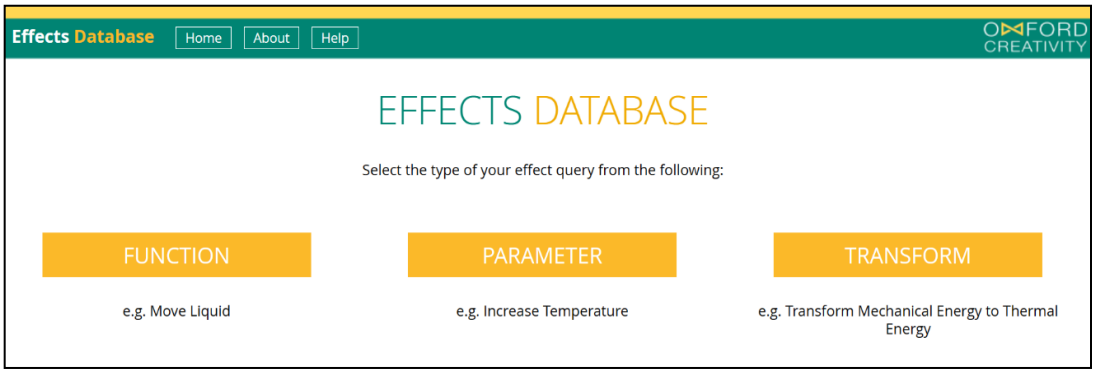
Entre as bases de efeitos físicos, destacam-se a *Oxford Creativity*[5] e a *Product Inspiration*[6]. Ambas, de acesso público e *online*, permitem estabelecer uma relação entre um efeito desejado e suas possíveis causas (Kraus et al., 2022). Nelas, o usuário especifica a função pretendida e obtém como resultado os fenômenos científicos que podem ser utilizados para cumpri-la (Makino et al., 2015).

A *Oxford Creativity*, fundada em 1998 por Karen Gadd, tem como objetivo tornar a metodologia TRIZ mais acessível por meio de sua interface digital, da variedade de variáveis disponíveis e da ampla cobertura de efeitos (Oxford Creativity, 2025a). A Figura 10 apresenta a interface da Base *Oxford Creativity*.

[5] http://wbam2244.dns-systems.net/EDB/
[6] https://www.productioninspiration.com/

Figura 10

Interface da base de efeitos físicos da *Oxford Creativity*

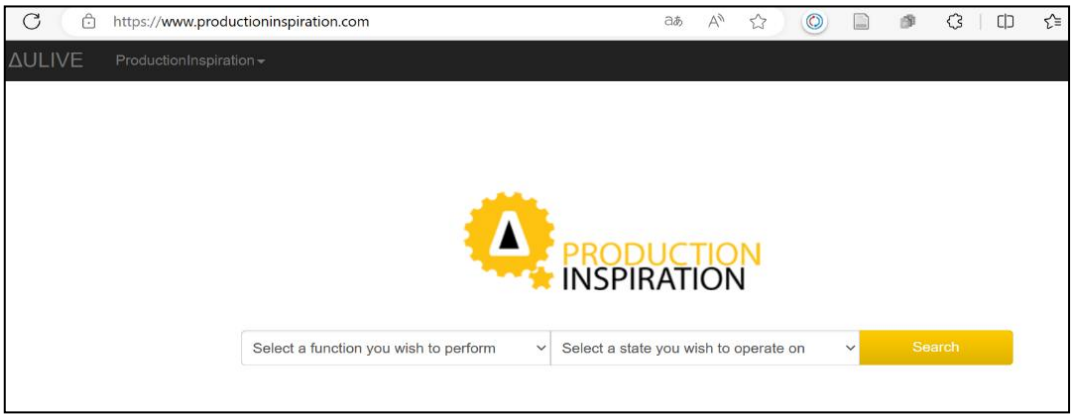


Nota. Obtido de Oxford Creativity (2025b).

A base de efeitos da empresa australiana Aulive – *Product Inspiration*[7] (Aulive, 2025) apresenta um número menor de efeitos em comparação à base *Oxford Creativity*, porém algumas de suas entradas são acompanhadas de animações gráficas. A Figura 11 apresenta a interface da base *Aulive – Product Inspiration*.

Figura 11

Interface da Base de efeitos físicos *Patent Inspiration*



Nota. Obtido de Aulive (2025).

Os efeitos físicos, sistematizados sob a forma de princípios inventivos, constituem padrões para resolver problemas inovadores em todas as áreas do conhecimento (Kim & Kim, 2012). Em domínios diferentes, o mesmo problema genérico é resolvido com o mesmo princípio inventivo (Ekmekci & Nebati, 2019; Lux, 2022). De acordo com a TRIZ, esses problemas técnicos podem favorecer o surgimento de novas tecnologias ou tendências de desenvolvimento tecnológico ou soluções técnicas diferentes (Li et al., 2023).

No âmbito da TRIZ, importantes constatações emergem:

(1) As inovações recorrem a efeitos científicos provenientes de campos distintos

[7] https://www.productioninspiration.com/

daquele ao qual pertence o produto ou serviço em desenvolvimento (Souili, Cavallucci, & Rousselot, 2015b);

(2) Problemas e soluções tendem a se repetir na indústria e na ciência (Souili, Cavallucci, & Rousselot, 2015b);

(3) Em domínios distintos, um mesmo problema genérico pode ser resolvido com o mesmo princípio inventivo, o que torna muitas invenções transferíveis de um domínio para outro (Ekmekci & Nebati, 2019; Lux, 2022);

(4) Soluções inventivas para um determinado problema são obtidas de forma sistemática, utilizando todo o potencial da ciência, da engenharia e de fora do campo do problema originalmente formulado (Chandra & Livotov, 2019).

Ao longo das décadas de 1960 e 1970, a TRIZ passou por sucessivos refinamentos conceituais e metodológicos. Contudo, em razão do isolamento político e científico do regime soviético, sua disseminação internacional foi limitada até o início da década de 1990, quando o colapso da União Soviética possibilitou a difusão da metodologia nos países ocidentais (Brad, 2023; Carvalho & Back, 2001). A partir desse período, a TRIZ passou a circular no cenário internacional, sendo progressivamente incorporada a ambientes acadêmicos e industriais[8].

Atualmente, a TRIZ é ensinada em diversas universidades ao redor do mundo e tem sido amplamente adotada por organizações globais como Hewlett-Packard, Ford Motors, Siemens, Boeing, IBM, Procter & Gamble, Mitsubishi, Xerox e Caterpillar, entre outras, com o objetivo de aprimorar processos de inovação e resolução sistemática de problemas (Brad, 2023; Ilevbare et al., 2013; Xu et al., 2019).

4.3 Procedimentos Metodológicos

Para o desenvolvimento do presente estudo, adotou-se uma abordagem de natureza descritiva e exploratória, com enfoque qualitativo, baseada em dados secundários obtidos por meio de uma Revisão Sistemática da Literatura (RSL). Essa revisão teve como objetivo analisar as tendências temporais e metodológicas das publicações que abordam a integração entre TRIZ e a mineração de textos de patentes, identificando os principais avanços e desafios da área. A RSL possibilita uma visão abrangente, permitindo identificar, analisar e sintetizar os estudos relevantes com o propósito de responder à questão de pesquisa proposta (Pradana et al., 2023).

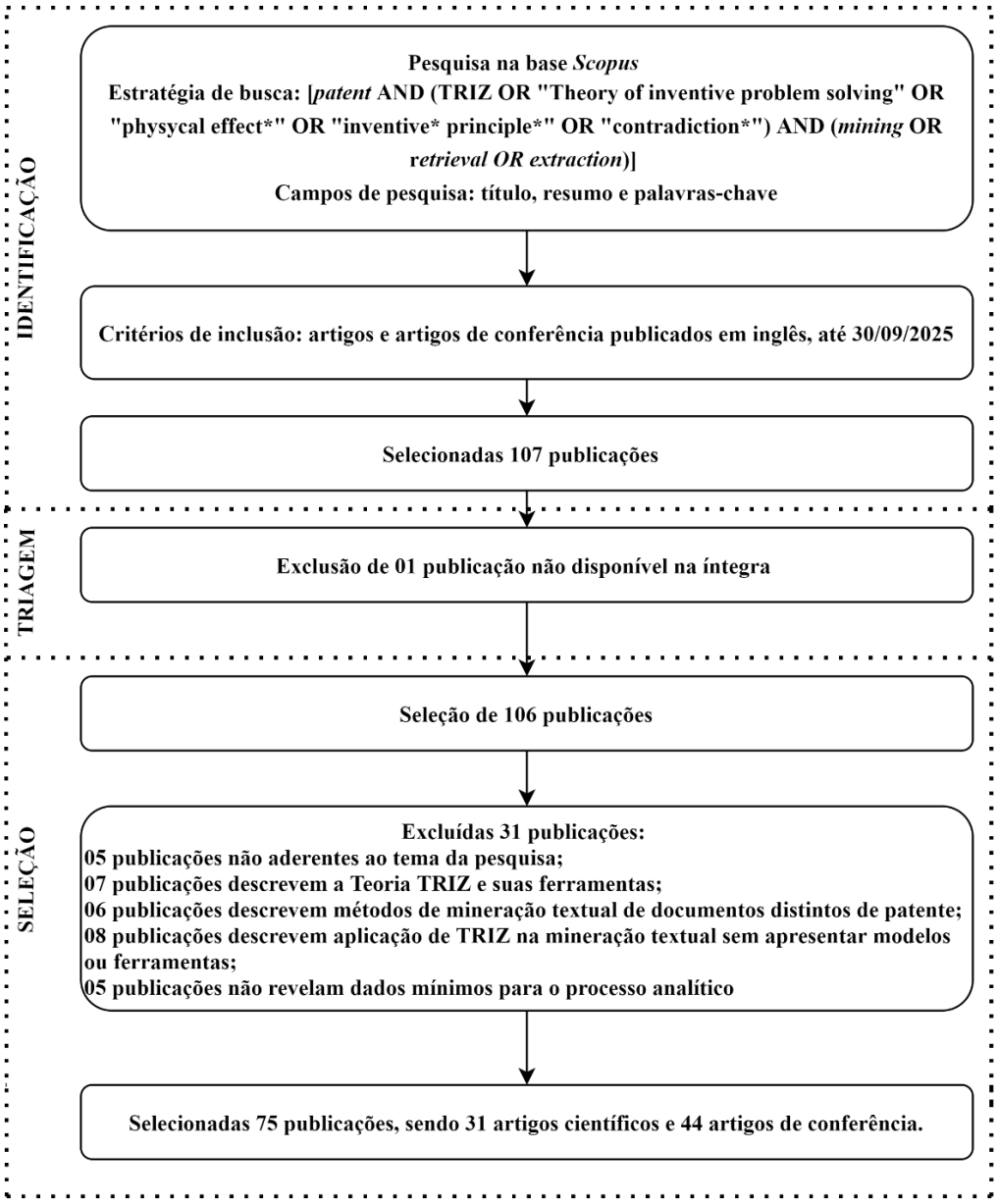
[8] Lista de Instituições acadêmicas: https://etria.eu/documents/TRIZ_academic_institutions.pdf. Acesso em 28, julho de 2025.

Para a condução da revisão sistemática, foi adotada a ferramenta PRISMA, com a definição de critérios de inclusão e exclusão, bem como o mapeamento da quantidade de registros. As informações são organizadas em um fluxo que facilita a interpretação dos resultados (*PRISMA*, 2025). A base de dados utilizada foi a *Scopus*, em virtude de sua natureza multidisciplinar e da disponibilização dos Anais da *International TRIZ Future Conference*, um importante canal de divulgação científica na área.

A estratégia de busca contempla termos relacionados à TRIZ e suas ferramentas e termos utilizados no contexto de mineração textual, resultando em uma expressão definida como [*patent AND (TRIZ OR "Theory of inventive problem solving" OR "physical effect*" OR "inventive* principle*" OR contradiction) AND (mining OR retrieval OR extraction)*]. Foram pesquisados os campos de título, resumo e palavras-chave, sendo selecionados artigos e artigos de conferência publicados em inglês até 30/09/2025. Na Figura 12 é apresentado o fluxograma da RSL.

Figura 12

Fluxograma da Revisão Sistemática da Literatura do Estudo 2



Nota. Adaptado de PRISMA (2025).

Na etapa de identificação, foram obtidas 107 publicações, das quais 106 artigos e artigos de conferência foram selecionados, tendo em vista a indisponibilidade do texto completo de uma das publicações. Após a leitura dos documentos, identificaram-se 31 artigos a serem excluídos por diversos motivos: cinco não apresentavam aderência ao tema da pesquisa; sete limitavam-se a descrever a TRIZ e suas ferramentas; seis tratavam de métodos de mineração textual aplicados a documentos que não eram patentes; oito relatavam a aplicação da TRIZ em mineração textual, mas sem apresentar modelos ou ferramentas; e cinco descreviam de forma genérica a ferramenta de mineração textual, sem fornecer dados mínimos necessários ao

processo analítico (Dybå et al., 2007).

Na etapa analítica, foram selecionadas 75 publicações, sendo 31 artigos científicos e 44 artigos de conferência. As publicações selecionadas na RSL foram examinadas quanto aos seguintes aspectos: a metodologia de extração de termos, incluindo as etapas de pré-processamento e análise de dados textuais; a ferramenta TRIZ empregada; os campos patentários explorados; os critérios adotados para avaliação dos métodos; as limitações dos estudos; e as propostas de pesquisas futuras. Para a análise de frequência de termos, foi utilizada a ferramenta *VOSviewer*. A relação completa das publicações selecionadas encontra-se no Apêndice D.

4.4 Resultados

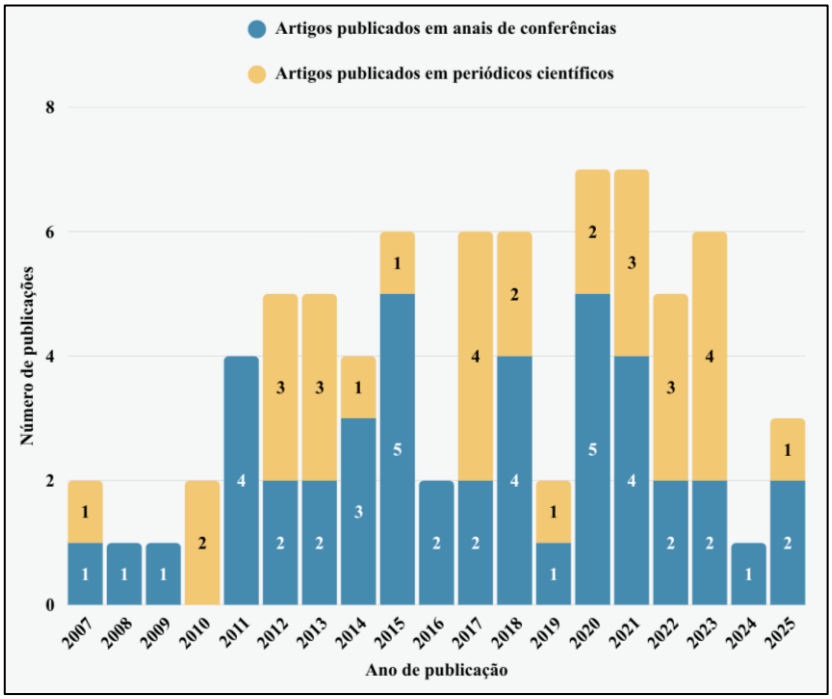
Esta seção detalha a análise das publicações selecionadas na RSL, estruturada em quatro subseções essenciais. Primeiramente, o perfil bibliográfico examinará a frequência anual, a origem geográfica (países e instituições) e os autores mais proeminentes, mapeando a evolução temporal e as áreas de concentração dos estudos. Em seguida, a segunda subseção apresentará um resumo das principais ferramentas TRIZ empregadas na mineração de textos de patentes. A terceira parte será dedicada à análise das tendências metodológicas identificadas no campo e, por fim, a última subseção sintetizará os desafios do campo que se apresentam para futuras pesquisas.

4.4.1 Resultados da Análise Bibliográfica

No que se refere à frequência de estudos sobre mineração de textos de patentes utilizando as ferramentas e conceitos da TRIZ, observa-se uma tendência ascendente a partir de 2005. Esse crescimento pode estar relacionado ao tempo necessário para que o mundo ocidental assimilasse e compreendesse a TRIZ e suas ferramentas, considerando que sua disseminação no Ocidente teve início em 1991. Além disso, a realização de conferências dedicadas ao tema, como a *International TRIZ Future Conference,* cuja primeira edição ocorreu em 2000, contribuiu para a ampliação do interesse e do reconhecimento da abordagem. A Figura 13 apresenta o gráfico da distribuição anual das publicações selecionadas na RSL.

Figura 13

Gráfico da distribuição anual das publicações selecionadas na Revisão Sistemática da Literatura do Estudo 2

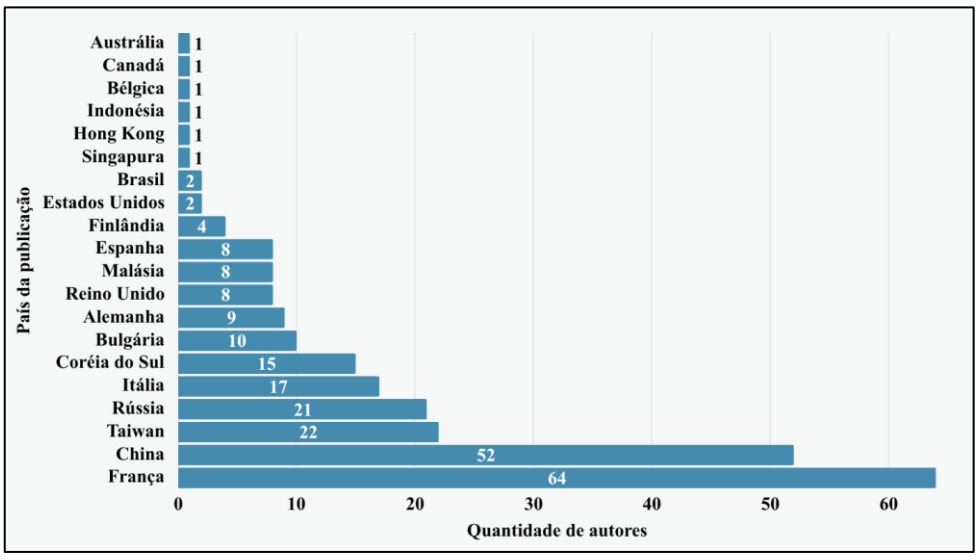


Nota. Dados da pesquisa (2025).

No cenário mundial, os estudos que envolvem a mineração textual de patentes e a aplicação da TRIZ destacam-se especialmente na França e na China. A Figura 14 apresenta a frequência de autores por país.

Figura 14

Gráfico da distribuição de autores por país



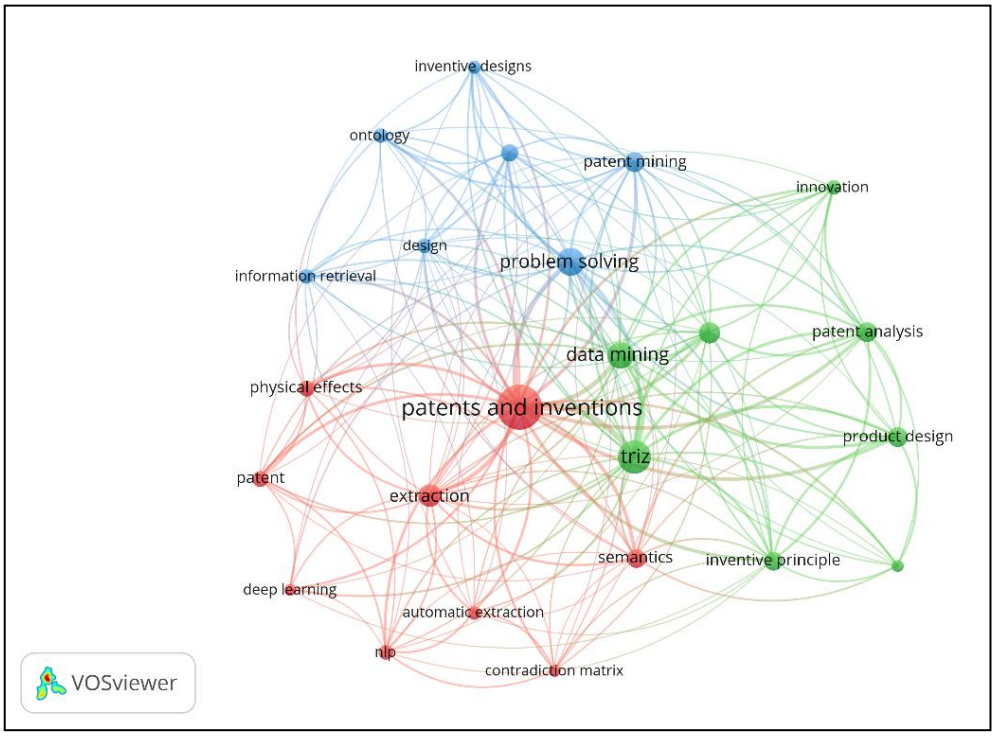
Nota. Dados da pesquisa (2025).
Nota. Duplicidade de dados para autores com múltiplas publicações.

O destaque dentre as instituições estrangeiras é o Instituto Nacional de Ciências Aplicadas (INSA) de Strasbourg, o Politécnico de Milão e o Laboratório de Inovação em Qualidade da Universidade de Bérgamo, que abrigam grupos de pesquisa sobre TRIZ (Brad, 2023). É possível identificar contribuições assíduas de alguns autores, como por exemplo Daria Berduygina (INSA/França), Denis Cavallucci (INSA/França), Gaetano Cascini (Universidade de Bérgamo/Itália) e Yanhong Lian (Universidade de Tecnologia de Hebei/China).

Para obter uma visão clara das principais áreas de concentração e da evolução temporal dos estudos selecionados na RSL, utilizou-se a ferramenta *VOSviewer*, que possibilitou uma análise descritiva e interpretativa com base na ocorrência e na inter-relação dos termos. Na análise de coocorrência de termos, o gráfico de rede resultante apresenta três agrupamentos (*clusters*) principais, diferenciados por cores (vermelho, verde e azul). Cada nó (círculo) representa um termo relevante identificado na literatura, e as ligações (linhas) indicam a frequência com que esses termos aparecem conjuntamente nos documentos. O tamanho dos círculos e rótulos reflete o peso ou importância do termo dentro do conjunto analisado — ou seja, quanto maior o círculo, maior a relevância do termo. Os agrupamentos temáticos da análise são demonstrados na Figura 15, por meio da rede bibliométrica de coocorrência de termos.

Figura 15

Rede bibliométrica baseada na coocorrência de termos



Nota. Dados da pesquisa (2025).

A rede evidencia três agrupamentos principais. O agrupamento vermelho, localizado no

centro da rede, representa o núcleo conceitual do mapa, tendo como termo central *"patents and inventions"*. Esse cluster conecta conceitos relacionados à extração automática e semântica, como *"semantics"*, *"extraction"*, *"deep learning"* e *"automatic extraction"*. Isso indica que uma parte significativa das pesquisas aborda o uso de métodos computacionais e de aprendizado profundo para analisar e extrair informações de patentes.

O agrupamento verde está associado ao uso da base de conhecimento TRIZ em contextos de mineração e análise de patentes. Os termos *"TRIZ"*, *"data mining"*, *"patent analysis"*, *"product design"* e *"inventive principle"* sugerem que esse conjunto de estudos explora a aplicação da TRIZ como ferramenta de inovação e resolução de problemas tecnológicos a partir de dados de patentes.

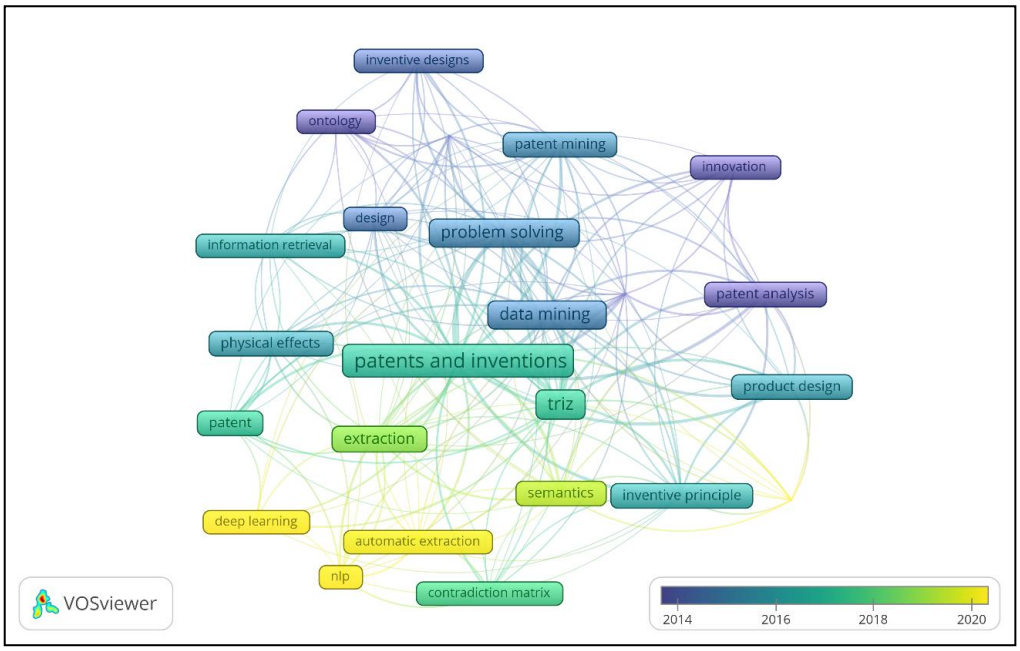
Por sua vez, o agrupamento azul concentra-se na recuperação e organização da informação, com termos como *"information retrieval"*, *"ontology"*, *"problem solving"*, *"patent mining"* e *"design"*. Esse cluster indica uma linha de pesquisa voltada à estruturação semântica e linguística das informações de patentes, enfatizando o acesso, a organização e a categorização do conhecimento tecnológico.

As múltiplas conexões entre os três agrupamentos evidenciam uma interdisciplinaridade significativa. O termo "data mining", por exemplo, aparece como ponto de interseção entre o cluster vermelho e o verde, sugerindo que técnicas de mineração de dados são aplicadas tanto para extração de informação quanto para análise baseada em TRIZ.

Com o propósito de analisar a progressão temporal dos termos, procedeu-se à realização de uma análise de sobreposição (*VOSviewer*), cujos resultados constam na Figura 16.

Figura 16

Análise de sobreposição de termos extraídos do corpus textual analisado



Nota. Dados da pesquisa (2025).

A Figura 16 evidencia que os termos "*patents and inventions*", "TRIZ", "*problem solving*", e "*data mining*" aparecem como nós centrais e de maior peso, o que indica sua alta frequência e relevância temática nas publicações analisadas. Esses termos formam o núcleo conceitual da rede, refletindo o foco predominante das pesquisas na aplicação de técnicas de mineração de dados e de textos em documentos de patentes com base nos princípios e ferramentas da TRIZ para aprimorar a resolução de problemas técnicos.

Ao redor desse núcleo, destacam-se subtemas complementares como "extraction", "*semantics*", "*ontology*", e "*information retrieval*", que representam os métodos e abordagens empregados na identificação e tratamento da informação técnica. Esses termos formam um segundo grupo de alta conectividade, associado ao uso de técnicas de PLN, modelagem semântica e análise ontológica no contexto de recuperação de inteligência técnica.

Em um nível mais periférico, surgem termos como "*deep learning*", "*automatic extraction*", "NLP", e "*contradiction matrix*", que indicam a incorporação de métodos mais recentes e automatizados, especialmente aprendizado de máquina e inteligência artificial, nas etapas de extração e classificação de informações em patentes. A coloração amarela desses termos, de acordo com a escala temporal apresentada (em torno de 2020), sugere que são tópicos emergentes e recentes na literatura, refletindo a tendência contemporânea de integrar TRIZ e técnicas avançadas de IA.

Por outro lado, termos com tonalidade azul, como "*inventive designs*", "*innovation*",

"*patent mining*" e "*product design*", indicam conceitos mais consolidados e presentes nos estudos anteriores (em torno de 2014–2016). Esses tópicos representam as bases iniciais da integração entre TRIZ e mineração de patentes, com foco em design inventivo e inovação tecnológica.

De modo geral, a estrutura da rede revela uma evolução temática: os estudos começaram centrados em design e inovação (2014–2016), avançaram para a mineração e análise semântica (2016–2018) e, mais recentemente, incorporaram aprendizado de máquina e PLN (a partir de 2020) como estratégias de extração e interpretação de conhecimento técnico a partir de patentes.

4.4.2 Aplicações das Ferramentas TRIZ na Mineração de Textos de Patentes

No contexto da mineração textual, a TRIZ fornece uma base conceitual e sistemática, por meio de um conjunto estruturado de ferramentas que aprimoram a extração e a análise da inteligência técnica. Entre as principais contribuições da TRIZ, destacam-se:

(a) a construção de consultas de pesquisa para explorar bancos de dados de patentes com base nos termos que caracterizam o problema, de acordo com o formalismo TRIZ (Becattini et al., 2015);

(b) a identificação de padrões e a extração de informações relevantes de textos de patentes (Berdyugina & Cavallucci, 2020b; Zhang et al., 2022), que podem ser aplicadas em diferentes domínios técnicos (Gongchang et al., 2014; Ni et al., 2020; Zhang et al., 2023) para resolver problemas, obter soluções inovadoras (Kim, Joung, et al., 2018; Zhang, Tan, et al., 2022; Zhang, Wang, et al., 2022) e compreender ou antecipar a evolução das tecnologias (Vicente Gomila & Palop Marro, 2013; Vicente-Gomila, 2014; Vicente-Gomila et al., 2017);

(c) a facilitação da classificação automática de patentes (Gongchang et al., 2014; Liang et al., 2008, 2009).

Assim, as ferramentas da TRIZ, quando combinadas à mineração de texto, permitem a extração de informações úteis de documentos de patentes, economizando tempo e esforço em análises manuais (Othman et al., 2017), além de possibilitar a filtragem de informações relevantes (Gongchang et al., 2014; Liang et al., 2008; Othman et al., 2017).

A Tabela 13 apresenta a síntese das principais ferramentas TRIZ utilizadas nos processos de mineração textual de patente e os respectivos estudos, com uma breve abordagem da metodologia.

Tabela 13

Ferramentas da TRIZ e estudos relacionados

Ferramenta TRIZ	Descrição e estudos relacionados
Parâmetros de Engenharia	Atributos que caracterizam o desempenho e a funcionalidade de um sistema, essenciais para a formulação de contradições (Berdyugina & Cavallucci, 2023). • Verhaegen et al. (2011) realizam extração de substantivos e adjetivos de títulos e resumos de patentes para identificar parâmetros físicos e requisitos funcionais. • Cavallucci et al. (2011b) definem parâmetros como propriedades que podem ser modificadas em um elemento técnico. • Trappey et al. (2013) ddescrevem metodologia de análise de similaridade de patentes baseada na extração de parâmetros e princípios inventivos. • Jiang et al.(2023) convertem informações de patentes em parâmetros TRIZ para análise cruzada. • Ding e Ma (2014) extraem substantivos, adjetivos e verbos para identificar parâmetros, propriedades e funções.
Princípios Inventivos	Conjunto de 40 princípios que auxiliam na superação de contradições de design e fomentam a inovação (Chandra & Livotov, 2019; Choi, Kang, et al., 2012; Liang et al., 2008) • Prickett e Aparicio (2012) desenvolvem ontologia do Sistema Técnico TRIZ para indexação por princípios inventivos. • Gongchang et al. (2014), Liang et al. (2008) e Liang et al. (2009) utilizam princípios para classificação automática de patentes. • Verhaegen et al. (2014) propõem algoritmo para classificar patentes segundo princípios inventivos. • Chandra Sekaran e Livotov (2019) enfatizam a análise interdisciplinar de princípios inventivos para melhorar os processos de inovação. • Kaliteevskii et al. (2020) aplicam princípios TRIZ para extração semântica automatizada.
Matriz de Contradições	Estrutura que relaciona parâmetros de engenharia conflitantes, resultando em contradições técnicas (Carvalho & Back, 2001; Prickett & Aparicio, 2012; Srinivasan & Kraslawski, 2006). • Liang e Tan (2007b) aplicam aprendizado de máquina para classificar patentes por contradições inventivas. • Cascini e Russo (2007) e Zhang et al. (2023) usam análise linguística para identificação de contradições. • Trappey et al. (2013) utilizam redes neurais para construir matrizes de contradições. • Guarino et al. (2020) e Guarino et al. (2022) desenvolvem abordagens de sumarização e modelos fundacionais (PatRiZ) para extração de parâmetros. • Montecchi e Russo (2015) esenvolvem ferramenta baseada no método *Function Oriented Search (FOS)* para identificação de contradições TRIZ. • Wang et al. (2016) propõem algoritmo de extração SAO para construir matriz de contradições de processos de domínio com base em padrões

Ferramenta TRIZ	Descrição e estudos relacionados
	semânticos. • Ding et al. (2017) utilizam a matriz de contradições TRIZ como estrutura para extração de conhecimento técnico de patentes. • Naveiro e de Oliveira (2018) aplicam a matriz Implantação da função de qualidade (QFD) para mapear contradições técnicas entre requisitos técnicos e de usuário. • Zhai et al. (2020) usam Doc2Vec para construir espaço semântico que aprimora o reconhecimento de contradições técnicas em patente. • Berduygina e Cavallucci (2020, 2021 e 2023) desenvolvem métodos linguísticos e estatísticos com TF-IDF, n-gramas e PLN para extração automática e agrupamento de parâmetros contraditórios em textos de patentes. • Trapp e Warschat (2025) aplicam *prompt engineering* com GPT-4 para extrair contradições da seção "Estado da arte" de patentes.
Efeitos Físicos	Fenômenos científicos aplicáveis à resolução de problemas inventivos no contexto TRIZ. • Russo e Montecchi (2011) identificam palavras-chave relacionadas a efeitos físicos. • Russo et al. (2012) propõem a ontologia Função–Comportamento–Estrutura de Efeitos Físicos (FBPHs) para classificar patentes com base em funções, comportamentos e efeitos físicos. • Korobkin et al. (2015, 2017) extraem descrições de efeitos físicos por análise estatística e semântica, utilizando LDA e árvores de dependência para marcação gramatical. • Fomenkova et al. (2017) utilizam ontologia para extrair efeitos físicos por análise semântica. • Valverde et al. (2017) desenvolvem método baseado em palavras-chave funcionais e bancos de dados de efeitos físicos para extração de informações de patentes. • Korobkin e Fomenkov (2018) e Korobkin et al. (2018, 2019) constroem matriz de funções físicas baseadas em efeitos extraídos de patentes. • Kaliteevskii et al. (2021) integram banco de dados de efeitos físicos para extrair conceitos semânticos considerando propriedades mecânicas, elétricas e outras, com ponderação por frequência de termos. • Chan et al. (2021) criam dicionário de efeitos físicos a partir de resumos de patentes do USPTO, estruturando-o em seções de assunto, objeto e ação. • Russo e Gervasoni (2022) desenvolvem modelo computacional que associa verbos funcionais a efeitos físicos ilustrados.
Padrões Evolutivos	Descrevem estágios previsíveis de desenvolvimento dos sistemas técnicos (Wang et al., 2010). • Wang et al. (2010) aplicam *KeyGraph* para mapear cenários tecnológicos e identificar padrões evolutivos em documentos de patente. • Jiang et al. (2025) identificam intensidades de tópicos e tendências de evolução de um campo do domínio.

Nota. Elaborado pela Autora (2025).

A análise dos estudos apresentados na Tabela 13 evidencia a diversidade de aplicações das ferramentas TRIZ na mineração de textos de patentes, abrangendo desde a identificação de parâmetros de engenharia e contradições técnicas até a extração de princípios inventivos e efeitos físicos. Essas abordagens demonstram que a integração entre TRIZ e técnicas computacionais requer não apenas a adaptação conceitual das ferramentas, mas também o uso de procedimentos robustos de processamento linguístico para garantir a precisão e a consistência dos resultados. Nesse contexto, observa-se que as etapas iniciais de tratamento textual assumem papel fundamental na qualidade da informação extraída, conforme discutido a seguir.

4.4.3 Processo de Mineração Textual de Patentes usando TRIZ

Para a mineração textual de patentes com o uso das ferramentas da TRIZ, adota-se o fluxo de trabalho convencional, que inclui uma etapa inicial de obtenção dos textos de patente, seguida pelo pré-processamento dos textos, a fim de prepará-los para o processo analítico subsequente.

Na etapa de obtenção dos textos, os estudos demonstram maior ênfase na exploração das seções de descrição e reivindicações dos documentos de patente, nas quais:

Berdyugina e Cavallucci (2020a), assim como Cascini e Russo (2007), argumentam que as reivindicações são mais ricas e precisas na citação de termos-chave, em comparação com os títulos e resumos. Além disso, Berdyugina e Cavallucci (2020a) destacam a importância das reivindicações como fonte primária de informações inventivas, propondo que a consideração de sua estrutura hierárquica pode aprimorar a qualidade da extração e reduzir o ruído informacional.

Apesar de sua relevância, observa-se uma menor disponibilidade das seções de reivindicações nos repositórios globais de patentes em relação aos resumos. Essa limitação pode representar um desafio metodológico na composição de bases de dados para o treinamento de algoritmos.

Por outro lado, Trapp et al. (2023) afirmam que a seção de antecedentes da invenção, presente no relatório descritivo da patente, tende a recuperar um maior número de soluções. Já Liang e Tan (2007b) sustentam que, em geral, os resumos e as descrições fornecem informações semânticas suficientes para a identificação dos Princípios Inventivos TRIZ empregados nas patentes.

Seguindo para a etapa de pré-processamento, que influencia diretamente os resultados das análises subsequentes (Souili, Cavallucci, & Rousselot, 2015b), as abordagens aplicadas

utilizam técnicas clássicas de PLN, como a eliminação de *stopwords* (Berdyugina & Cavallucci, 2020a), a marcação de classe gramatical (Berdyugina & Cavallucci, 2020a, 2021), a lematização — processo de redução das formas flexionadas de uma palavra à sua raiz lexical (Kaliteevskii et al., 2020, 2021; Souili, Cavallucci, & Rousselot, 2015b), a tokenização (Kaliteevskii et al., 2021; Souili, Cavallucci, & Rousselot, 2015b; Verhaegen et al., 2014) e a extração de palavras recorrentes, porém sem conteúdo técnico (Berdyugina & Cavallucci, 2021).

Alguns estudos também incorporam a segmentação de frases (Berdyugina & Cavallucci, 2020a; Huang et al., 2023) ou a segmentação em nível de palavras (Liang & Tan, 2007), o que pode facilitar etapas posteriores de processamento e análise.

Para a extração de parâmetros, princípios inventivos, efeitos físicos e funções a partir de textos de patentes, combinam-se técnicas de análise semântica com métodos estatísticos. Utilizam-se ferramentas voltadas ao PLN baseado em dados linguísticos, como a biblioteca *Python Natural Language Toolkit* (NLTK) (Kaliteevskii et al., 2020, 2021) e o banco de dados lexical *WordNet* (Liang et al., 2008; Russo, 2011; Yoon & Kim, 2012). Além disso, aplicam-se métodos de aprendizado de máquina, como Doc2Vec (Chan et al., 2021; Kaliteevskii et al., 2020, 2021; Zhai et al., 2020), *K-means* (Cao et al., 2016; Kaliteevskii et al., 2020, 2021) e modelagem de tópicos (*Latent Dirichlet Allocation – LDA*) (Berdyugina & Cavallucci, 2021; Kaliteevskii et al., 2020), voltados ao agrupamento e à análise semântica dos textos.

No que se refere às técnicas de análise semântica, estas possibilitam uma compreensão mais profunda do conteúdo textual, indo além da mera correspondência de palavras para revelar significados e relações conceituais. Diversos estudos demonstram que tais abordagens reduzem significativamente o tempo de processamento em comparação com métodos manuais (Spreafico & Spreafico, 2021).

A abordagem SAO é amplamente utilizada para a extração de relações entre entidades (Choi et al., 2012; Korobkin & Fomenkov, 2018; Othman et al., 2017; Park, Ree, et al., 2013; Park, 2012; Vicente-Gomila et al., 2017) e para a identificação de conceitos e relacionamentos tecnológicostecnológicos (Park, Kim, et al., 2013). No entanto, Vicente-Gomila (2014) ressalta que nem todas as relações SAO extraídas são significativas no contexto da literatura técnica, o que torna indispensável uma avaliação especializada.

A pesquisa orientada a funções (*Function-Oriented Search* – FOS) tem sido discutida como alternativa para aprimorar a análise de patentes (Montecchi & Russo, 2015; Zhang et al., 2023). Contudo, sua eficácia é limitada por ambiguidades semânticas decorrentes do uso de termos tecnológicos (Choi et al., 2012). Para mitigar essa limitação, Liu e Li et al. (2020)

propõem a extração de informações funcionais, termos técnicos e códigos da IPC, de modo a classificar informações de função e selecionar patentes com maior potencial para oferecer soluções relevantes.

Por fim, os estudos apontam o uso de ontologias, que podem ser estruturadas de diferentes formas:

(a) orientadas a fatos, modelando informações funcionais da tecnologia (Choi et al., 2012);
(b) baseadas em Função, Comportamento e Estrutura (Russo, 2011);
(c) voltadas a domínios específicos, identificando sentenças formadas por verbos e substantivos (Spreafico & Spreafico, 2021);
(d) genéricas, como o Método de *Design* Inventivo (IDM), que permite extrair automaticamente três conceitos fundamentais para a formulação de problemas e soluções técnicas: problemas, soluções parciais e parâmetros (Souili, Cavallucci, & Rousselot, 2015c, 2015a, 2015b; Souili, Cavallucci, Rousselot, et al., 2015).

A partir de 2018, observa-se o desenvolvimento de métodos de mineração textual com base na TRIZ que aplicam IA para aprimorar diversas tarefas, entre as quais se destacam:

(a) a análise e classificação de textos (Huang et al., 2023; Jiang et al., 2023; Joseph et al., 2016; Souili, Cavallucci, & Rousselot, 2015a, 2015b);
(b) a recuperação de frases que contenham a principal contradição de uma patente, permitindo observar o contexto semântico (Guarino et al., 2021, 2024; Guarino, Samet, & Cavallucci, 2020);
(c) a extração de contradições TRIZ a partir de textos de patentes utilizando o GPT-4, da OpenAI (Trapp & Warschat, 2025);
(d) a identificação de padrões textuais que denotam funções (em verbos), propriedades (em adjetivos) e contexto (em substantivos) (Dewulf & Childs, 2023);
(e) a classificação de soluções inventivas em diferentes domínios, a partir da identificação de um problema-alvo, com o uso de modelos baseados em redes neurais (Ni et al., 2021).
(f) a identificação de recursos para apoiar o processo de inovação de produtos, obtendo hotspots de pesquisa por meio do treinamento do ChatGPT com o suporte de ferramentas de IA e TRIZ (Du et al., 2025).

Nas últimas duas décadas, os avanços substanciais na análise de patentes (Aristodemou & Tietze, 2018), especialmente com a aplicação de inteligência artificial e redes neurais

artificiais, têm possibilitado progressos significativos. Ainda assim, persistem desafios relevantes, que serão discutidos na seção seguinte.

4.4.4 Desafios da Integração de TRIZ com Mineração Textual

Os estudos identificam diversos desafios na integração das técnicas de mineração de texto com as ferramentas TRIZ, os quais podem ser agrupados em quatro categorias principais:

(a) **Estrutura, morfologia e sintaxe dos documentos patentários** – Os textos de patentes apresentam grande variação terminológica, incluindo sinônimos e termos ambíguos que assumem significados distintos em diferentes contextos técnicos, gerando confusão semântica (Berdyugina & Cavallucci, 2021; Gongchang et al., 2014; Liang & Tan, 2007). A extensão dos documentos e a complexidade estrutural das frases (Kang et al., 2018) resultam em dados ruidosos, exigindo significativo esforço humano para análise (Liang & Tan, 2007). Além disso, a terminologia heterogênea dificulta a identificação de vínculos semânticos, podendo levar a extrações imprecisas (Cascini & Zini, 2011; Fomenkova et al., 2017).

(b) **Complexidade e atualização das ferramentas TRIZ** – O uso prático das ferramentas TRIZ ainda é percebido como complexo (Berdyugina & Cavallucci, 2023; Kaliteevskii et al., 2020; Souili, Cavallucci, & Rousselot, 2015b). Muitas delas permanecem excessivamente abstratas e genéricas (Chan et al., 2021; Choi et al., 2012), e carecem de ontologias formalizadas (Berdyugina & Cavallucci, 2023; Souili & Cavallucci, 2013). Ademais, os termos originais empregados nas ferramentas TRIZ encontram-se desatualizados frente às soluções tecnológicas emergentes descritas nas patentes recentes (Berdyugina & Cavallucci, 2022a; Liang et al., 2008; Trapp & Warschat, 2025).

(c) **Adequação linguística das abordagens de mineração textual** – A maioria das abordagens analisadas foi desenvolvida para o idioma inglês, o que restringe sua aplicabilidade a patentes redigidas em outros idiomas e evidencia limitações no processamento multilinguístico.

(d) **Disponibilidade e qualidade dos textos de patentes** – A digitalização incompleta ou inadequada dos documentos ainda impede a execução de análises automáticas (Cavallucci et al., 2011b). Com frequência, apenas os resumos são disponibilizados em formatos adequados para a extração automática de informações, o que limita o acesso ao conteúdo técnico mais detalhado e

prejudica o processamento da linguagem natural e da semântica latente (Aristodemou et al., 2017).

Apesar desses desafios, que delineiam uma agenda de pesquisa complexa e interdisciplinar, os avanços recentes demonstram um potencial promissor para a integração entre IA, mineração textual e TRIZ. Em um cenário em que o acesso à informação se torna um diferencial competitivo, o aprimoramento dessas técnicas pode transformar significativamente o modo como o conhecimento inventivo é extraído, analisado e aplicado na inovação tecnológica.

4.5 Discussão

Os métodos de mineração e recuperação de informações aplicados à análise de conteúdo textual têm evoluído de forma significativa nas últimas décadas, impulsionados pelo avanço de algoritmos, técnicas de processamento de linguagem natural (PLN) e métodos baseados em inteligência artificial. No contexto das patentes, essa evolução está diretamente associada ao crescimento exponencial do número de documentos disponíveis em bases públicas, o que ampliou substancialmente o acesso a informações estratégicas para a proteção da propriedade intelectual e o apoio à inovação tecnológica. Apesar desse avanço, a literatura indica que a pesquisa voltada especificamente à recuperação e à análise automatizada de documentos de patente ainda se encontra em processo de consolidação (Liu, Li, et al., 2020).

A integração entre TRIZ e mineração textual tem promovido avanços relevantes na extração de inteligência técnica. Essa integração possibilita a análise sistemática de soluções inventivas, a identificação de contradições técnicas e a exploração de novas aplicações tecnológicas, ampliando o potencial inovador dos documentos de patente.

Os resultados da RSL evidenciam que os conceitos fundamentais da TRIZ — problemas, soluções, parâmetros e contradições — oferecem uma estrutura conceitual robusta e compatível com aplicações em linguística computacional e sistemas de inteligência artificial. Essa estrutura contribui para a organização do conhecimento técnico extraído e para o apoio à tomada de decisão, ao desenvolvimento de produtos e à inovação tecnológica.

No que se refere às agendas futuras de pesquisa, os estudos analisados indicam a necessidade de aperfeiçoar o uso de ferramentas de PLN voltadas especificamente à extração de informações inventivas (Berdyugina & Cavallucci, 2021; Choi, Kang, et al., 2012; Li & Tate, 2010). Em termos morfológicos e sintáticos, há forte incentivo à ampliação de investigações

sobre correlações semânticas, com foco em sinônimos, antônimos, homônimos, parônimos e fenômenos de polissemia, recorrentes nos textos de patentes (Berdyugina & Cavallucci, 2021; Kaliteevskii et al., 2020).

Quanto às ferramentas TRIZ, destaca-se a necessidade de atualização e refinamento terminológico, de modo a acompanhar a rápida evolução das tecnologias descritas nas patentes mais recentes (Berdyugina & Cavallucci, 2022a; Liang et al., 2008; Trapp & Warschat, 2025). Já no campo da mineração textual, os estudos sugerem comparar sistematicamente a qualidade e a relevância das informações extraídas em diferentes seções dos documentos de patente, como resumo, descrição e reivindicações, a fim de identificar quais partes oferecem dados mais consistentes para a análise e a geração de soluções inovadoras.

4.6 Considerações Finais

Este estudo oferece uma contribuição relevante ao campo da extração de inteligência técnica a partir de patentes, ao apresentar uma visão abrangente das metodologias existentes e dos desafios associados à integração entre TRIZ e mineração textual. A análise realizada demonstra que, embora as técnicas de mineração de texto tenham avançado de forma significativa, sua aplicação a documentos de patente ainda exige adaptações linguísticas, estruturais e terminológicas específicas, capazes de lidar com a complexidade sintática e semântica desses textos.

A integração entre TRIZ e mineração textual mostrou-se particularmente relevante por fornecer uma estrutura conceitual sistemática para organizar o conhecimento extraído — incluindo problemas, soluções, parâmetros, contradições e efeitos físicos. Sob a perspectiva da Visão Baseada no Conhecimento (KBV), esses achados evidenciam que a mineração textual atua como um mecanismo habilitador da capacidade absortiva organizacional, entendida como a habilidade de reconhecer o valor do conhecimento externo, adquiri-lo, assimilá-lo, transformá-lo e explorá-lo com fins estratégicos (Cohen & Levinthal, 1990; Zahra & George, 2002).

Os resultados indicam avanços claros nas quatro dimensões da capacidade absortiva: (i) aquisição, por meio do acesso automatizado a grandes volumes de patentes; (ii) assimilação, via técnicas semânticas, ontológicas e linguísticas que reduzem ambiguidade e ruído informacional; (iii) transformação, ao articular o conhecimento extraído com estruturas como TRIZ, matrizes de contradição e ontologias; e (iv) exploração, ao apoiar diretamente a inovação,

o design de produtos, a resolução de problemas técnicos e a antecipação de tendências tecnológicas.

A análise metodológica permitiu identificar lacunas importantes na literatura, especialmente no que se refere à replicabilidade dos estudos, à profundidade analítica das abordagens e à adaptação das técnicas de mineração textual a diferentes idiomas — aspecto particularmente relevante para a língua portuguesa. Ainda assim, os resultados confirmam que a combinação entre TRIZ e mineração de textos de patentes constitui uma área de pesquisa promissora, caracterizada pela crescente adoção de técnicas de aprendizado de máquina e aprendizado profundo para explorar, de forma mais eficiente, o maior repositório mundial de informações tecnológicas.

Do ponto de vista prático, destaca-se o valor da inteligência técnica extraída de patentes no contexto empresarial, especialmente para pequenas e médias empresas, ao subsidiar a identificação de oportunidades de mercado e apoiar decisões estratégicas em ambientes altamente competitivos e dinâmicos.

Quanto às limitações do estudo, reconhece-se que a estratégia de busca pode ter empregado termos excessivamente amplos ou restritivos, o que pode ter resultado na exclusão de trabalhos relevantes (Shaheen et al., 2023). Além disso, a exclusão de publicações que não apresentavam aplicações explícitas de TRIZ à mineração textual pode ter limitado a abrangência dos achados. Ainda assim, os desafios identificados foram sistematizados em categorias estruturais, terminológicas, linguísticas e tecnológicas, oferecendo um panorama consolidado das barreiras que ainda restringem a integração plena entre TRIZ e mineração textual. Em síntese, os resultados reforçam o elevado potencial dessa intersecção como suporte estratégico à inovação tecnológica, à tomada de decisão e ao uso competitivo das patentes.

5 ESTUDO 3: DESENVOLVIMENTO DE UMA ONTOLOGIA BASEADA NA TEORIA DA SOLUÇÃO INVENTIVA DE PROBLEMAS (TRIZ)

Resumo

As patentes constituem uma fonte valiosa de inteligência técnica, oferecendo um repositório de conhecimento que permanece subutilizado na academia, nos negócios e na indústria, contextos em que tais informações poderiam gerar benefícios significativos. A ampliação do acesso a bases de dados de patentes tem o potencial de fornecer *insights* de ponta em diferentes domínios tecnológicos. No campo da linguística computacional, a Teoria da Resolução Inventiva de Problemas (TRIZ), desenvolvida a partir de extensivas análises de patentes, representa uma oportunidade promissora para o avanço da mineração textual aplicada a documentos escritos em português. Este estudo tem como objetivo desenvolver uma ontologia fundamentada em conceitos extraídos da engenharia e do conhecimento científico, contemplando tanto efeitos físicos puros quanto aplicados, capazes de sugerir potenciais soluções genéricas. A rede semântica proposta, construída a partir de termos oriundos do domínio das patentes, busca aprimorar os processos de recuperação da inteligência técnica embutida nesses documentos. Classificado como uma pesquisa aplicada de caráter exploratório, o trabalho introduz uma ontologia composta por termos hierarquicamente organizados e inter-relacionados em língua portuguesa. Essa estrutura estabelece uma ponte lexical e semântica, possibilitando a aplicação de ferramentas computacionais para a mineração de inteligência técnica em patentes redigidas em português.

Palavras-chave: Patente. TRIZ. Ontologia.

Abstract

Patents represent a rich and underutilized source of technical intelligence, encompassing a vast repository of knowledge with significant potential for application across academia, industry, and business. Expanding access to patent databases can unlock valuable insights and foster innovation across a wide range of technological domains. Within the field of computational linguistics, the Theory of Inventive Problem Solving (TRIZ, originally developed through systematic analysis of patents, offers a robust conceptual foundation for advancing text mining methodologies, particularly for documents written in Portuguese. This study aims to develop an ontology grounded in engineering and scientific concepts, integrating both pure and applied physical effects to support the identification of potential generic solutions. The proposed semantic network, derived from terminology specific to the patent domain, seeks to enhance the retrieval and utilization of technical intelligence embedded in patent texts. As an applied exploratory effort, this research introduces an ontology comprising hierarchically organized and semantically interrelated terms in Portuguese. This lexical and semantic framework provides a bridge for the effective use of computational tools in mining and leveraging technical intelligence from Portuguese-language patents.

Keywords: *Patent. TRIZ. Ontology.*

5.1 Introdução

As patentes, enquanto fonte primária de inteligência técnica (Liwei, 2022; Xu et al., 2022), oferecem informações tecnológicas essenciais que apoiam o avanço da ciência, da tecnologia e da inovação, além de contribuírem para a gestão estratégica empresarial (Belenzon, 2012; Han et al., 2006). Em última análise, ampliam o potencial inovador das economias modernas e fortalecem a pesquisa científica e tecnológica.

Entretanto, o volume massivo de documentos de patentes, que supera a capacidade de análise exclusivamente humana, demanda o uso de ferramentas de recuperação da informação para garantir análises eficazes (Krestel, Chikkamath, et al., 2021; Tseng et al., 2007). Nesse contexto, a recuperação automatizada de informações técnicas torna-se um elemento crucial na análise de patentes. A literatura especializada descreve diversos métodos de mineração textual, os quais são identificados e analisados nos Estudos 1 e 2 desta tese. Mais recentemente, observa-se a expansão do uso de técnicas de IA, que têm contribuído para aumentar a precisão e a relevância dos resultados obtidos (Mandl, 2009).

As estruturas textuais singulares e a semântica especializada presentes nos documentos de patentes impõem desafios específicos aos sistemas de recuperação da informação (Fall et al., 2003). Isso ocorre porque tais documentos combinam campos estruturados (semanticamente consistentes e formatados), com campos não estruturados, como reivindicações, resumos e descrições técnicas (Zanella et al., 2023). Nessas seções, é comum a presença de termos associados a função, efeito e propósito, além de terminologia técnica e variações sinônimas. Essa abordagem baseada na relação problema–função constitui um princípio central da TRIZ, a qual identifica padrões fundamentais de resolução de problemas por meio da análise sistemática de patentes em diversos domínios científicos e de engenharia (Cong & Tong, 2008). Tal situação foi constatada no Estudo 1 e, por esse motivo, no Estudo 2 é aprofundada a análise de métodos de mineração de inteligência técnica em documentos de patentes fundamentados na TRIZ.

A partir da TRIZ, diversas pesquisas têm se concentrado no desenvolvimento e aplicação de ontologias para viabilizar processos de recuperação de conhecimento a partir das informações contidas em milhões de invenções tecnológicas. Em geral, essas abordagens envolvem a mineração de textos de patentes utilizando efeitos físicos descritos pela teoria. Em muitos casos, ontologias são construídas a partir de elementos extraídos automaticamente das

funções sintáticas dos textos (An et al., 2021; Teng et al., 2024). Outras recorrem a ontologias linguísticas, como a *WordNet* (Wu et al., 2010), que exigem a tradução dos textos para o inglês, ou a ontologias específicas de domínio, adaptadas a áreas técnicas particulares (Sarica et al., 2020; Spreafico & Spreafico, 2021; Trappey et al., 2024; Vincent & Cavallucci, 2018).

O presente estudo tem como objetivo construir uma ontologia de domínio que integre os efeitos físicos da TRIZ, adaptada ao contexto das patentes redigidas em língua portuguesa. A ontologia foi projetada para aplicação em ferramentas de mineração de texto e contém léxicos aplicáveis a distintos campos tecnológicos. Essa estrutura possibilita a expansão vocabular por meio de enriquecimento semântico, favorecendo a integração com ontologias específicas já existentes.

A questão central que orienta esta pesquisa é: Como desenvolver uma ontologia de domínio baseada nos efeitos físicos da TRIZ que apoie, de forma semântica e linguística, a mineração de textos de patentes em português? Para responder a essa pergunta, o estudo adota uma metodologia de natureza aplicada e exploratória, que reaproveita recursos de conhecimento previamente disponíveis como base para a construção do modelo conceitual da ontologia proposta.

O estudo está estruturado em cinco seções: (i) revisão da literatura, que aborda os métodos de aprendizado de máquina, o uso de ontologias para mitigar a escassez de dados anotados e as contribuições da TRIZ; (ii) seção metodológica, na qual são detalhados os procedimentos adotados na concepção da ontologia de domínio; (iii) apresentação dos resultados; (iv) discussão; e, por fim, (v) considerações finais.

5.2 Revisão da Literatura

Na seção de revisão da literatura, são apresentadas as abordagens de aprendizado de máquina voltadas à extração de inteligência técnica em textos de patente, com ênfase nas dificuldades estruturais e semânticas desses documentos. Também são discutidas a escassez de dados anotados, que limita a eficiência da mineração textual, e as contribuições da TRIZ, à luz do conhecimento acumulado nos Estudos 1 e 2 desta tese.

5.2.1 Abordagens de Aprendizado de Máquina para a Extração de Inteligência Técnica de Documentos de Patente

A pesquisa em recuperação de informações de patentes é relativamente recente,

concentrando-se no desenvolvimento de técnicas e métodos capazes de recuperar, de forma eficaz e eficiente, documentos relevantes em resposta a solicitações de busca específicas (Shalaby & Zadrozny, 2019). Em termos gerais, a análise de patentes pode ser conduzida a partir de duas abordagens principais: (i) análise bibliométrica**,** baseada em campos estruturados dos documentos, como classes tecnológicas, datas de concessão e citações (Giordano et al., 2023; Miric et al., 2023); e (ii) mineração de texto, que explora informações não estruturadas presentes no corpo principal dos documentos, como descrições, reivindicações e resumos (Abbas et al., 2014; Aristodemou & Tietze, 2018; Lupu, 2017; Miric et al., 2023). No contexto da mineração de texto, surgem desafios específicos que demandam não apenas a adaptação de métodos de Recuperação de Informação (RI) e IA, mas também o desenvolvimento de novas abordagens mais adequadas às particularidades do domínio patentário (Krestel et al., 2022).

Sistemas de recuperação de informação aprimorados por IA têm o potencial de transformar significativamente a pesquisa e a análise de patentes. Entretanto, para o treinamento eficaz desses modelos, são necessárias milhões de amostras devidamente anotadas (Singh, 2018). Um dos principais entraves nesse campo é a escassez de conjuntos de dados anotados com alta qualidade (Xu et al., 2019). A rotulagem, ou anotação de dados, constitui uma etapa crucial do pré-processamento, na qual os dados brutos são manualmente enriquecidos com rótulos que conferem o contexto necessário para que modelos de aprendizado de máquina realizem análises precisas (Singh, 2018).

Nos documentos de patentes, a estrutura textual e a semântica específicas, especialmente a combinação entre terminologia jurídica e técnica, impõem desafios significativos às aplicações de PLN (Berdyugina & Cavallucci, 2020a). O estilo linguístico e o vocabulário especializado influenciam fortemente a eficácia dos sistemas de mineração textual aplicados à análise patentária. Esses sistemas frequentemente enfrentam dificuldades para identificar conexões semânticas entre termos ou expressões distintas que se referem ao mesmo componente ou função (Cascini & Zini, 2011). Além disso, a linguagem altamente técnica das patentes dificulta a extração automática de termos-chave (Yue, Liu, Hou, et al., 2023). A redundância de expressões, a presença de termos polissêmicos (com múltiplos significados dependentes do contexto) e o uso de sinônimos entre diferentes domínios técnicos tornam o processo de extração ainda mais complexo (Hu et al., 2018).

Para extrair informações de textos de patentes, os métodos de mineração textual geralmente se baseiam em três paradigmas de aprendizado: não supervisionado, que dispensa dados rotulados; supervisionado, que depende de dados previamente rotulados; e semissupervisionado, no qual um classificador é treinado com dados anotados e,

posteriormente, utilizado para rotular progressivamente os dados não anotados (Choi et al., 2021). No entanto, a rotulagem de dados é um processo custoso, demorado e suscetível a erros (Hu et al., 2018), além de exigir especialistas com conhecimento técnico aprofundado (Choi et al., 2021). Já o aprendizado não supervisionado, embora menos oneroso, não considera as nuances semânticas do texto, o que limita sua capacidade de capturar significados contextuais mais complexos (Salton & Buckley, 1988).

No domínio das patentes, a escassez de dados de treinamento anotados (Blume et al., 2024; Trapp & Warschat, 2025) representa um obstáculo significativo para o desenvolvimento de modelos supervisionados robustos em tarefas de mineração textual. Essa limitação restringe a capacidade de generalização dos algoritmos e compromete a precisão na identificação de padrões técnicos relevantes. Nesse contexto, as ontologias emergem como uma alternativa promissora, capaz de suprir lacunas semânticas e favorecendo a extração de informações em cenários de baixa disponibilidade de dados anotados.

5.2.2 Ontologias: Alternativa Promissora na Mineração Textual de Patentes

No contexto da mineração textual de patentes, as ontologias de domínio têm sido amplamente empregadas como estruturas de metadados capazes de classificar termos ou objetos e explicitar suas inter-relações (Prickett & Aparicio, 2012). Conforme definido por Cavallucci et al. (2011a), ontologias constituem uma representação formalizada do conhecimento em um domínio, construída a partir de uma conceituação ou perspectiva particular, com o objetivo de facilitar o compartilhamento e a comunicação de informações. Assim, configuram-se como uma alternativa promissora para superar as limitações dos métodos exclusivamente supervisionados ou não supervisionados.

Nos Estudos 1 e 2 desta teses, constatou-se que diversos trabalhos utilizam ontologias em métodos de mineração textual de patentes, funcionando como pontes para a identificação de problemas técnicos e soluções potenciais em diferentes campos (Prickett & Aparicio, 2012; Soo et al., 2005; Souili, Cavallucci, & Rousselot, 2015b; Taduri et al., 2019). Além disso, observa-se a combinação de ontologias com técnicas de PLN, como análise sintática, para extrair informações significativas dos textos (Jing et al., 2023). Esse uso integrado contribui para mitigar inconsistências terminológicas, ampliar a interpretação semântica de palavras-chave e fornecer maior contexto situacional, promovendo também interoperabilidade entre distintas fontes de informação (Prickett & Aparicio, 2012).

Nesse sentido, uma das aplicações mais relevantes das ontologias no domínio de patentes está associada à Teoria da Resolução Inventiva de Problemas (TRIZ). Diversos estudos

traçam uma analogia entre os textos de patentes, nos quais soluções inventivas para problemas técnicos são descritas, e os recursos baseados na TRIZ, que utilizam princípios inventivos para prever efeitos aplicáveis à resolução de problemas (Cong & Tong, 2008; Prickett & Aparicio, 2012). As coleções de efeitos físicos da TRIZ reúnem conceitos oriundos da engenharia e do conhecimento científico, com ênfase em sua aplicação prática para a solução de problemas (Ilevbare et al., 2013). Essa abordagem possibilita a realização de buscas orientadas pela função desejada, favorecendo a recuperação de fenômenos científicos aplicáveis a diferentes campos da ciência e da engenharia (Makino et al., 2015). A Tabela 14 sintetiza os principais estudos que descrevem ontologias derivadas da TRIZ aplicadas à mineração de textos de patentes.

Tabela 14

Síntese dos principais estudos que descrevem ontologias derivadas de TRIZ para mineração de texto de patentes

Autores /ano	Estrutura da ontologia	Limitações observadas	Principais contribuições
Prickett e Aparicio (2012)	A ontologia é composta por quatro classes principais: sistema técnico TRIZ (intervalo numérico entre 0 e 1 que estabelece um padrão de evolução tecnológica), recursos (funcionais, ambientais e de sistema), função e princípios inventivos TRIZ.	A maximização da reutilização requer conhecimento prévio das ferramentas TRIZ, mas aspectos de clareza e legibilidade ainda são pouco explorados na definição do índice atribuído ao sistema técnico.	A estrutura de classes de funções desenvolvida permite analisar diversas fontes de conhecimento e know-how de projeto, além de suportar a interoperabilidade com ontologias semelhantes, utilizando o banco de dados de funções do NIST (Instituto Nacional de Padrões e Tecnologia, EUA).
Russo (2011)	Baseia-se em dois modelos principais: o Elemento–Nome da Propriedade–Valor da Propriedade, derivado do modelo SAO, e o modelo Estrutura–Função–Comportamento, que envolve Função, Efeito Físico/Químico e Parâmetro de Projeto.	A inclusão de efeitos químicos pode demandar a incorporação de termos específicos de domínio, indicando a necessidade de integração com bases de conhecimento especializadas (como classificadores de patentes) em pesquisas futuras.	Recomenda a exploração de aspectos linguísticos (relações lexicais e semânticas) por meio de ferramentas avançadas, como dicionários, tesauros conceituais, navegadores linguísticos e catálogos de verbos funcionais (ex.: NIST). Também enfatiza a integração dessas ferramentas com ontologias de projeto conceitual (como FBS e

Autores /ano	Estrutura da ontologia	Limitações observadas	Principais contribuições
			ENV) e propõe a extração de palavras-chave de classificações de patentes para o desenvolvimento de tesauros.
Souili, Cavallucci e Rousselot (2015b)	A ontologia do Método de *Design* Inventivo (IDM) utiliza conceitos e relacionamentos para organizar e facilitar a extração de conhecimento, estruturando-se em torno de classes de problemas, soluções parciais e contradições, incorporando elementos, parâmetros e valores.	Necessita de um *corpus* de treinamento com marcadores linguísticos bem definidos para apoiar a identificação do conhecimento relacionado ao IDM.	Propõe uma ontologia genérica e aplicável a diversas áreas do conhecimento, possibilitando a recuperação de termos polissêmicos, cujos significados variam conforme o contexto de uso.
Vincent e Cavallucci (2018)	A Ontologia Formal Básica é estruturada em duas classificações principais: continuantes (entidades que persistem no tempo) e ocorrências (eventos que se manifestam no tempo e no espaço). Os 39 Parâmetros de Engenharia são descritores de objetos e categorizados como continuantes dependentes.	O estudo aplica ontologias derivadas da TRIZ a um domínio específico, limitando sua abrangência.	Os resultados indicam a necessidade de ampliar o conjunto de dados para incluir mais continuantes biológicos. Alguns parâmetros da TRIZ foram adaptados em relação às versões originais, visando maior relevância para a biologia.

Nota. Elaborado pela Autora (2025).

Os estudos demonstram que os parâmetros físicos e os princípios da TRIZ constituem recursos valiosos para a extração de inteligência técnica a partir de documentos de patente. Nesse contexto, têm-se intensificado os esforços para expandir o corpus terminológico por meio da integração com outras bases de conhecimento, com o objetivo de atualizar e enriquecer os termos disponíveis.

Entre as bases de conhecimento integradas às ontologias para fins de expansão vocabular, destacam-se os sistemas de classificação de patentes (Phan et al., 2018), a *Wikipedia* (Chao et al., 2021) e ontologias pré-existentes voltadas a campos específicos do conhecimento (Taduri et al., 2019).

No entanto, a diversidade linguística impõe desafios significativos às tarefas de mineração textual e construção de ontologias, bem como à adequação das metodologias e

ferramentas da TRIZ à língua portuguesa. Como a maior parte desses recursos está originalmente disponível em inglês (Zaniro et al., 2024), torna-se necessário realizar adaptações estruturais ao traduzi-los ou aplicá-los ao português, em virtude das diferenças entre classes gramaticais e construções sintáticas. requererendo adaptações estruturais quando transpostos para o português, em razão das diferenças entre classes gramaticais e construções sintáticas.

À luz das limitações observadas na literatura e das particularidades linguísticas do português, definiu-se uma abordagem metodológica voltada à construção de uma ontologia capaz de integrar os efeitos físicos da TRIZ ao domínio das patentes. A seguir, são apresentados os procedimentos adotados para o desenvolvimento do modelo conceitual e sua aplicação no contexto da mineração de textos.

5.3 Procedimentos Metodológicos

Este estudo é classificado como uma pesquisa aplicada de natureza exploratória, voltada a detalhar as condições em investigação (Wang et al., 2022). Para a construção do modelo conceitual da ontologia de domínio, foi adaptado o Banco de Dados Multilíngue de Efeitos Físicos, proposto por Zaniro et al. (2024) de modo a alinhar-se ao objetivo do estudo (Suárez-Figueroa et al., 2009), isto é, desenvolver uma ontologia semântica funcional baseada em efeitos físicos da TRIZ para a extração de recursos representativos em documentos de patentes. Nesse contexto, o Banco de Dados Multilíngue de Efeitos Físicos oferece uma lista de sugestões de efeitos associados a funções e objetos, cujas relações são validadas por especialistas. Contudo, para dar suporte à arquitetura de software necessária à implementação de processos de aquisição de conhecimento, esses termos precisam ser organizados hierarquicamente e inter-relacionados, de forma a possibilitar o treinamento de modelos voltados à mineração de inteligência técnica em patentes.

Para o desenvolvimento da ontologia foi utilizada a metodologia NeOn (Baonza, 2010), com os requisitos definidos por meio de um Documento de Especificação de Requisitos de Ontologia (*Ontology Requirements Specification Document* - ORSD) (Suárez-Figueroa et al., 2009). Esse documento estabelece o objetivo principal da ontologia, sua cobertura, granularidade prevista, linguagem de programação, público-alvo, usos pretendidos, bem como os requisitos funcionais e não funcionais. Os requisitos não funcionais “referem-se às características, qualidades ou aspectos gerais não relacionados ao conteúdo da ontologia que a ontologia deve satisfazer” (Angeloni, 2003. p.973). Já os requisitos funcionais “referem-se ao

conhecimento específico a ser representado pela ontologia" (Angeloni, 2003, p. 973).

O desenvolvimento da ontologia compreende três fases:

(1) Fase conceitual: envolve a aquisição de conhecimento, a análise dos dados de entrada da ontologia candidata (neste caso, a ontologia a ser reutilizada), a definição dos requisitos por meio do ORSD e o planejamento da iteração.

(2) Fase de *design*: concentra-se na especificação dos requisitos da ontologia. Nessa etapa, a arquitetura e o modelo conceitual foram definidos com base nos requisitos estabelecidos. Os termos coletados na fase conceitual passaram por curadoria manual, sendo normalizados e traduzidos, com preservação das classes e dos relacionamentos da ontologia candidata.

(3) Fase de implementação: consiste na conversão da conceituação desenvolvida em um modelo formal para um código legível por máquina, com o objetivo de facilitar a recuperação de inteligência técnica em documentos de patentes.

Para a construção da rede semântica foi utilizada a ferramenta de código aberto Protégé[9]. As classes, subclasses e os relacionamentos entre elas (McGuinness & van Harmelen, 2009) foram definidos na Linguagem de Ontologia da Web (*Ontology Web Language* - OWL) (Jarrar, 2002), que permite estruturar, definir e instanciar ontologias.

5.4 Resultados

A seção de resultados apresenta o processo de estruturação da base multilíngue de efeitos TRIZ que serviu de base para a construção do modelo conceitual da ontologia, descrita na segunda subseção.

5.4.1 Estruturação da Base Multilíngue de Efeitos Físicos TRIZ

Os bancos de efeitos físicos da TRIZ constituem ferramentas *online* que permitem aos usuários selecionar uma função técnica (ou ação) em combinação com um tipo de objeto (como sólido, líquido ou gás) sobre o qual a ação será aplicada. Essa combinação gera uma lista de efeitos físicos sugeridos, relevantes para a tarefa especificada (Gao & Zhu, 2015). Esses efeitos podem ser aplicados inclusive na solução de problemas fora do domínio original, e múltiplos efeitos podem ser combinados em cenários complexos, ampliando as possibilidades de resolução de problemas de maior sofisticação (Navas, 2013).

[9] http://protege.stanford.edu

Entre os bancos de dados TRIZ disponíveis ao público, destacam-se o *Oxford Creativity*[10] e o *Product Inspiration*[11], ambos de acesso aberto e caracterizados por não exigirem conhecimento prévio sobre os fundamentos da TRIZ. Dessa forma, configuram-se como ferramentas acessíveis e de apoio à inovação, especialmente úteis para usuários interessados em resolução criativa de problemas.

A base *Oxford Creativity*, fundada em 1998 por Karen Gadd, tem como principal objetivo tornar o método TRIZ mais acessível, tanto por meio de sua representação digital quanto pela abrangência das variáveis e efeitos contemplados.

A base de efeitos *Product Inspiration*, desenvolvida pela empresa australiana Aulive, disponibiliza um conjunto mais restrito de efeitos em comparação à base *Oxford Creativity*. Contudo, distingue-se por apresentar recursos visuais adicionais, incluindo animações gráficas que ilustram a aplicação prática dos efeitos em diferentes contextos.

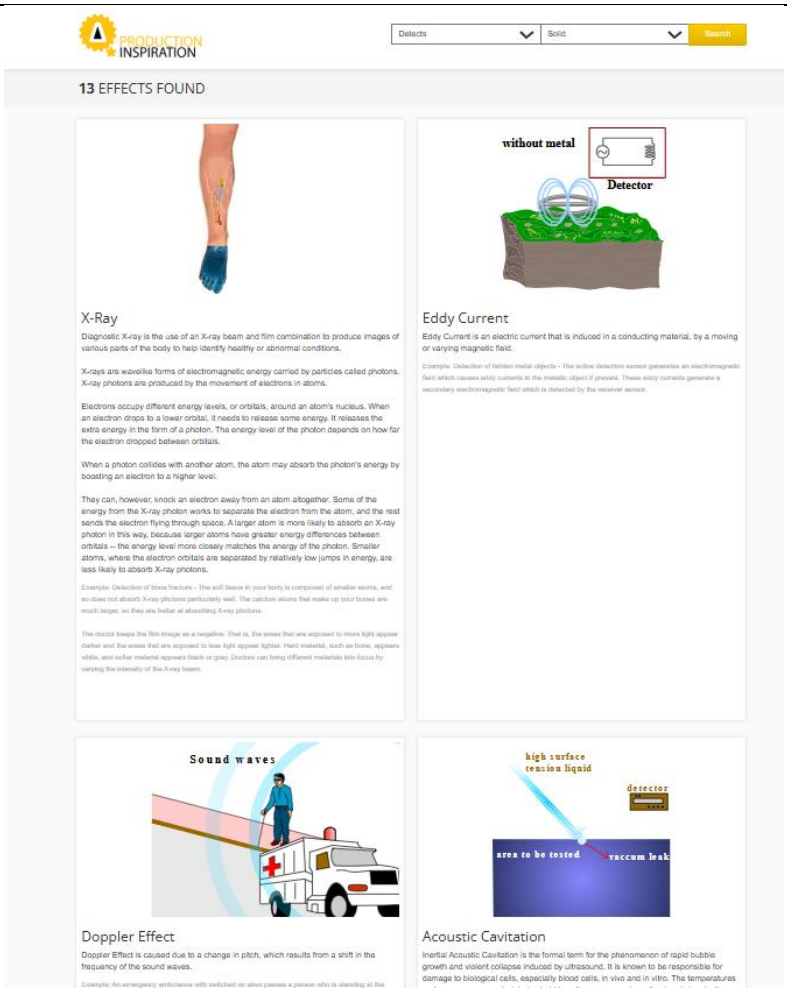
Como exemplo, apresenta-se o resultado de uma busca realizada em ambas as bases por meio de uma pesquisa "por função", utilizando o termo "*detection*" no campo *task* (ação) e "*solid*" no campo *target* (objeto). Os resultados obtidos são apresentados na Figura 17.

Figura 17

Interface do resultado da consulta nas bases *Oxford Creativity* e *Product Inspiration*

Oxford Creativity

[10] https://wbam2244.dns-systems.net/EDB/
[11] https://www.productioninspiration.com/

Product Inspiration	

Nota. Obtido de Oxford Creativity (2025b) e Aulive (2025).

Na comparação entre as bases de efeitos, a *Oxford Creativity* apresenta um número significativamente maior de efeitos técnicos para a mesma função e objeto do que a *Product Inspiration* (154 contra 13 resultados).

A base TRIZ multilíngue desenvolvida por Zaniro et al. (2024) foi construída a partir da coleta de dados das duas fontes por meio de *web scraping*: 23.681 entradas foram extraídas da *Oxford Creativity* e 1.166 da *Product Inspiration*, sendo 496 delas duplicadas.

Os dados foram pré-processados e categorizados nos campos *type* (efeito ou aplicação), *task* (ação), *target* (objeto) e *mode* (função, parâmetro ou transformação), totalizando 24.351 registros de efeitos físicos, todos em inglês.

Para a adaptação à ontologia, foram identificadas 670 entradas incompletas da *Product Inspiration*, contendo apenas os campos *type* e *mode*. Essas entradas passaram por curadoria manual para adequação gramatical e posterior tradução para o português. Em seguida, toda a base foi traduzida, com atenção às particularidades semânticas, como a distinção entre verbos, substantivos e adjetivos.

Por fim, os registros classificados como transformações ou parâmetros foram removidos, resultando em 11.196 entradas, focadas na identificação de relações entre tarefas e objetos que geram soluções técnicas imediatas e conclusivas.

5.4.2 Construção da Ontologia

Para a definição dos requisitos da ontologia, adotou-se o roteiro proposto no Documento de Especificação de Requisitos de Ontologia (ORSD – *Ontology Requirements Specification Document*), que orienta o processo de identificação, organização e formalização dos elementos conceituais essenciais ao modelo ontológico. A Tabela 15 apresenta a descrição das tarefas executadas e o desenvolvimento para a construção do ORSD.

Tabela 15

Síntese das tarefas para a construção do Documento de Especificação de Requisitos de Ontologia (ORSD)

Descrição da tarefa	Desenvolvimento
Determinar o propósito principal da ontologia	Criação de um modelo de conhecimento baseado na terminologia de efeitos físicos da TRIZ para permitir a recuperação de inteligência técnica de patentes.
Determinar o escopo	Efeitos físicos derivados de conceitos dentro do corpo de conhecimento científico e de engenharia, aplicados à resolução de problemas.
Determinar a linguagem de implementação	OWL (Linguagem de Ontologia da Web)
Identificar o usuário final pretendido	Estudantes – Construir o arcabouço teórico para seus estudos. Empreendedores/Inventores – Identificar lacunas tecnológicas ou aspectos específicos de campos tecnológicos. Pesquisadores – Analisar e detalhar o campo tecnológico sob investigação. Setor industrial – Identificar oportunidades tecnológicas e monitorar tendências de mercado. Governo/Instituições públicas – obtenção de dados para estabelecimento de políticas públicas.
Identificar o uso pretendido	Recuperação de inteligência técnica de patentes – Apoiar atividades de pesquisa em diversos campos do conhecimento. Apoio à pesquisa de ponta – Avaliar o grau de inovação de uma determinada tecnologia.
Identificar requisitos da ontologia	
Requisitos não funcionais	Termos em português; Termos nomeados e categorizados como Tarefa, Objeto e Efeito Físico.
Requisitos funcionais	Termos categorizados em subclasses, apresentados em classes gramaticais específicas. A subclasse Tarefa sempre compreende um verbo.

Descrição da tarefa	Desenvolvimento
	As subclasses Objeto e Efeito Físico compreendem um substantivo. A subclasse Objeto definida genericamente, com instâncias: sólido, sólido dividido, líquido, gasoso e área.
Pré-glossário de termos	Tarefa: Representa uma ação potencial voltada para o desenvolvimento. Objeto: Definido como o meio pelo qual uma tarefa atinge seu propósito. Efeito Físico: É o resultado de uma Tarefa em um Objeto, que pode ser um fenômeno científico (Efeito puro) ou um Efeito aplicado.

Nota. Adaptado de Baonza (2010).

O conjunto de requisitos foi validado por três especialistas em patentes, com amplo conhecimento tanto do conjunto de dados quanto das ferramentas baseadas na metodologia TRIZ. A partir desses requisitos, a fase de *design* desenvolveu o modelo conceitual da ontologia, estabelecendo os relacionamentos entre as classes Tarefa, Objeto e Efeitos Físicos.

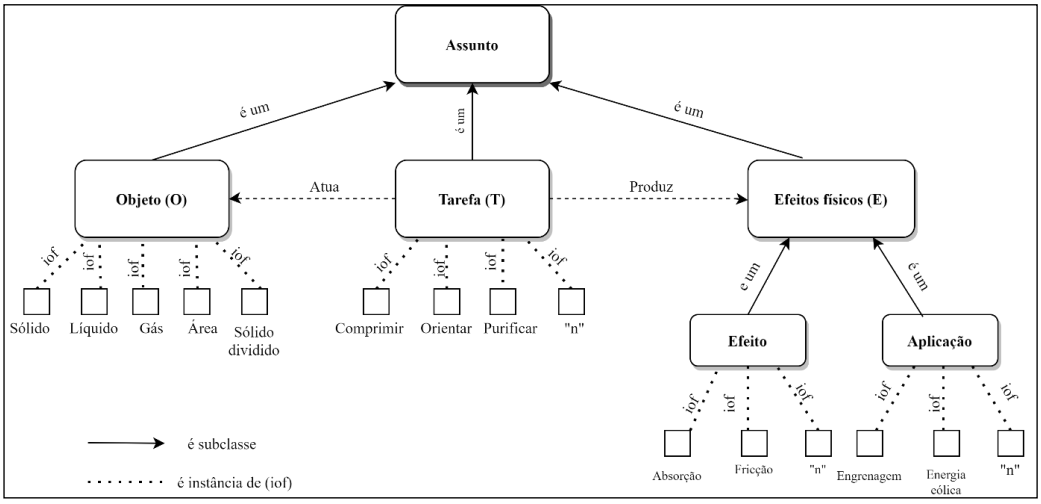
A representação conceitual introduz uma classe geral denominada Assunto, que se relaciona diretamente com suas subclasses Tarefa, Objeto e Efeitos Físicos, representando as principais categorias do domínio. As instâncias da ontologia, correspondentes ao nível mais granular de representação, foram definidas com base nas potenciais aplicações do modelo.

A subclasse Efeitos Físicos é dividida em duas subclasses: Efeito puro, que abrange fenômenos de natureza estritamente científica, e Efeito aplicado. Dessa forma, qualquer termo classificado como instância de Efeito puro ou Efeito aplicado é, por definição, também uma instância de Efeitos Físicos.

A Figura 18 apresenta a representação do projeto conceitual da ontologia, oferecendo uma visão geral das classes, subclasses e instâncias.

Figura 18

Representação do projeto conceitual da ontologia



Nota. O rótulo “n” indica a existência de termos adicionais que não são exibidos explicitamente na figura.

A ontologia está disponível em repositório público [https://osf.io/cbaef/], contendo 11.196 entradas, cada uma representada como um relacionamento ternário que integra as subclasses Tarefa, Objeto e Efeitos Físicos.

Na fase de implementação, a ontologia será testada com o objetivo de extrair soluções genéricas (conceituais) a partir de documentos de patentes. Nesse contexto, ela funciona como uma rede semântica abrangente de conceitos de engenharia, estruturada por associações semanticamente significativas, contribuindo para o aprimoramento da qualidade das informações recuperadas.

Quando essas informações são combinadas com a intervenção humana, que envolve experiência, interpretação e reflexão (Jarrar, 2002), a ontologia se torna um facilitador para a geração de soluções específicas e factuais, aplicáveis a problemas concretos.

A Tabela 16 apresenta um exemplo de relacionamento definido na ontologia proposta, ilustrando a integração entre as subclasses e a estrutura conceitual que sustenta o modelo.

Tabela 16

Exemplo de relacionamento ternário da ontologia

Objeto ← *atua*	Tarefa → *produz*	Efeito Físico	Tipo de Efeito
Gás	Absorver	Absorção	Puro
Gás	Absorver	Carvão ativado	Aplicado
Líquido	Absorver	Limpeza	Aplicado
Líquido	Absorver	Hidrogenação	Puro
Gás	Absorver	Absorção	Puro

Nota. Dados da pesquisa (2025).

A partir de uma função ou tarefa técnica associada a um tipo de objeto, genericamente definido como sólido, sólido dividido, área, líquido ou gasoso, ao qual a ação (Tarefa) será aplicada, são sugeridos efeitos físicos relevantes para a tarefa especificada. Esses efeitos, derivados do conhecimento científico e de engenharia e validados por especialistas, são caracterizados como potenciais soluções genéricas voltadas à execução da função pretendida.

A expansão terminológica por meio de sinônimos, hiperônimos e hipônimos deve ser investigada com o propósito de ampliar a recuperação de informações em diferentes domínios técnicos.

5.5 Discussão

A ontologia proposta reduz a dependência de especialistas para tarefas de classificação,

sendo projetada para a mineração de texto em múltiplas áreas do conhecimento, incorporando indicadores linguísticos genéricos relacionados a propriedades tecnológicas. Os termos são extraídos com base em similaridades lexicais e semânticas, podendo ser enriquecidos por meio da incorporação de termos presentes nos textos analisados e da associação com outras ontologias. Ainda assim, o processo pode demandar revisão manual para a eliminação de termos com ruído.

A sistematização de termos relacionados a efeitos físicos provenientes de diversas disciplinas científicas organiza o conhecimento em classes, subclasses e instâncias de maneira estruturada e semanticamente coerente. Essa organização assegura a consistência lexical e semântica em relação à terminologia empregada em documentos de patentes, promovendo alinhamento conceitual entre o modelo ontológico e os textos patentários.

As instâncias associadas às subclasses funcionam como marcadores gramaticais exploráveis por ferramentas de PLN, viabilizando a extração de termos semanticamente relacionados e fortalecendo o potencial da ontologia como instrumento de mineração semântica e recuperação de inteligência técnica em documentos de patente.

Ao integrar uma terminologia geral da ciência e da tecnologia com um conjunto estruturado de termos técnicos que representam ações aplicadas a objetos, a ontologia contribui para mitigar ambiguidades linguísticas nas etapas de análise sintática e semântica dos textos. Essa integração resulta em maior precisão na seleção de documentos e em recomendações mais relevantes de soluções técnicas aplicáveis a diferentes campos tecnológicos.

Embora a lista de termos ainda demande expansões e atualizações contínuas, a força da base de conhecimento está na amplitude conceitual e na aderência ao vocabulário técnico característico das patentes, o que confere robustez e aplicabilidade prática à proposta. Essa abordagem também reduz a dependência de grandes volumes de dados de treinamento, comum em métodos de aprendizado de máquina, tornando os processos de mineração de texto mais acessíveis, eficientes e transparentes.

A abrangente cobertura de termos vinculados a efeitos físicos, aliada ao potencial de integração com outras bases de conhecimento, possibilita inferir ações e relações técnicas implícitas, mesmo quando estas não estão explicitamente descritas nas patentes. Essa capacidade inferencial amplia o escopo e a qualidade da recuperação de informações, reduz o risco de omissão de termos relevantes e eleva o nível de abrangência semântica do sistema.

Adicionalmente, a ontologia representa uma contribuição inédita ao processamento de patentes em língua portuguesa, oferecendo suporte à mineração textual nesse idioma. Tanto sua estrutura conceitual quanto a metodologia adotada configuram um avanço pioneiro na

exploração de conteúdos tecnológicos em português, com potencial de aplicação em contextos multilíngues e interdisciplinares.

5.6 Considerações Finais

A contribuição teórica deste estudo concentra-se no desenvolvimento conceitual de uma ontologia semântica funcional baseada nos efeitos físicos da TRIZ, concebida como uma base semântica e linguística de apoio à mineração de textos de patentes. A ontologia proposta contempla instâncias que atuam como marcadores gramaticais, permitindo a identificação e análise de termos semanticamente relacionados que, considerando as particularidades do discurso patentário, sinalizam soluções técnicas.

Do ponto de vista prático, a estrutura ontológica viabiliza o uso de ferramentas computacionais para aquisição e organização de inteligência técnica a partir de patentes, fornecendo suporte direto às atividades de pesquisa, desenvolvimento e inovação (PD&I). A base de conhecimento, composta por termos validados por especialistas e extraídos da análise de milhares de invenções, apresenta alta correspondência com o léxico técnico real, reforçando sua aplicabilidade em contextos industriais e acadêmicos.

Em termos gerenciais, a ontologia constitui um instrumento estratégico de gestão do conhecimento, capaz de apoiar processos de prospecção tecnológica, vigilância de patentes e identificação de oportunidades de inovação, contribuindo para o fortalecimento do ecossistema de inovação orientado por dados.

Como limitação, ressalta-se a necessidade de expansão contínua da ontologia, tanto em abrangência lexical quanto em diversidade de fontes, a fim de ampliar sua cobertura semântica. Ademais, sua avaliação empírica em contextos multilíngues e interdisciplinares ainda requer investigação aprofundada.

No que se refere às perspectivas futuras, prevê-se a integração da ontologia com bases complementares de conhecimento, como tesauros técnicos, dicionários de sinônimos e antônimos e classificadores internacionais de patentes, além do enriquecimento do modelo por meio da associação a ontologias específicas de domínio. Propõe-se ainda sua aplicação em modelos híbridos de PLN e IA visando aprimorar a precisão, a eficiência e a interpretabilidade dos processos de recuperação de informação tecnológica.

6 ESTUDO 4: DESENVOLVIMENTO DE UM MÉTODO DE MINERAÇÃO DE INTELIGÊNCIA TÉCNICA EM PATENTES UTILIZANDO UMA ONTOLOGIA BASEADA NOS EFEITOS FÍSICOS DA TRIZ

Resumo

A mineração textual de patentes constitui uma área de investigação relevante em razão do grande volume de informação técnica não estruturada presente nesses documentos. Sob a perspectiva da Visão Baseada no Conhecimento (KBV), as patentes representam ativos de conhecimento codificado cujo valor estratégico depende da capacidade organizacional de reconhecer, assimilar e aplicar o conhecimento nelas contido. Contudo, a elevada variação terminológica dos textos patentários, marcada pelo uso de termos funcionais, ambíguos e dependentes do domínio tecnológico, impõe desafios à identificação de conceitos tecnológicos-chave. Nesse contexto, a TRIZ, fundamentada em parâmetros, propriedades, funções e contradições, apresenta elevada aderência à lógica descritiva das patentes, ao oferecer uma estrutura conceitual sistemática para a representação de problemas e soluções técnicas. Sua integração à mineração textual contribui para a redução de ambiguidades semânticas e para a organização do conhecimento extraído, favorecendo a transformação da informação técnica em conhecimento estruturado e acionável. Dessa forma, a articulação entre TRIZ e mineração textual fortalece dimensões centrais da capacidade absortiva, especialmente a assimilação e a transformação do conhecimento externo, ampliando o potencial estratégico das patentes como fonte de inovação. Nesse cenário, este estudo teve como objetivo desenvolver e validar um método para a mineração de inteligência técnica de patentes, com foco na extração do relacionamento semântico ternário Tarefa–Objeto–Efeito Físico (T-O-EF), utilizando uma Ontologia TRIZ como base estruturada. A metodologia adotou uma abordagem exploratória qualitativa e uma arquitetura híbrida que integra LLM para extração inicial e um componente de Lógica de Decisão Derivativa. O método foi validado por sete especialistas, alcançando desempenho global de 73,26%, superando o modelo original da ontologia e validando um modelo escalável para aquisição automatizada de conhecimento e geração de sugestões conceituais baseadas na TRIZ.

Palavras-chave: Mineração textual. Patente. LLM. TRIZ.

Abstract

Patent text mining is a relevant and timely research area due to the large volume of unstructured technical information contained in these documents. From the perspective of the Knowledge-Based View (KBV), patents represent codified knowledge assets whose strategic value depends on an organization's ability to recognize, assimilate, and apply the knowledge they contain. However, the high terminological variability of patent texts—characterized by the use of functional, ambiguous, and domain-dependent terms—poses significant challenges to the identification of key technological concepts. In this context, TRIZ, grounded in parameters, properties, functions, and contradictions, shows strong alignment with the descriptive logic of patent documents by providing a systematic conceptual structure for representing technical problems and solutions. Its integration with text mining contributes to reducing semantic ambiguity and organizing extracted knowledge, thereby facilitating the transformation of technical information into structured and actionable knowledge. As a result, the articulation between TRIZ and text mining strengthens core dimensions of absorptive capacity, particularly

the assimilation and transformation of external knowledge, enhancing the strategic potential of patents as a source of innovation. Within this framework, this study aimed to develop and validate a method for mining technical intelligence from patents, focusing on the extraction of the ternary semantic relationship Task–Object–Physical Effect (T–O–PE) using a TRIZ-based ontology as a structured foundation. The methodology adopted a qualitative exploratory approach and a hybrid architecture integrating large language models (LLMs) for initial extraction with a Derivative Decision Logic component. The method was validated by seven experts, achieving an overall performance of 73.26%, outperforming the original ontology model and confirming a scalable approach for automated knowledge acquisition and the generation of TRIZ-based conceptual suggestions.

Keywords: *Textual mining. Patent. LLM. TRIZ.*

6.1 Introdução

Com base em evidências empíricas apresentadas nos Estudos 1 e 2 desta tese, observa-se que as ferramentas de Processamento de Linguagem Natural (PLN) desempenham papel fundamental na extração, análise e representação de informações provenientes de grandes volumes de dados textuais. Essas ferramentas não apenas ampliam a capacidade de compreensão e comunicação científica, como também viabilizam o tratamento de conteúdos cuja complexidade e escala seriam impraticáveis sem o suporte da Inteligência Artificial (IA) (Mandl, 2009). Sob a perspectiva da Visão Baseada no Conhecimento (KBV), tais tecnologias atuam como mecanismos habilitadores da conversão de informação técnica codificada em conhecimento organizacional estratégico, ao reduzir barreiras cognitivas associadas ao uso de conhecimento externo.

No contexto da aquisição de inteligência técnica, especialmente em domínios intensivos em conhecimento, as tecnologias de mineração textual associadas à IA emergem como alternativas metodológicas robustas e escaláveis. O emprego dessas abordagens permite a automação de tarefas cognitivamente exigentes, como reconhecimento de padrões, extração de conceitos e identificação de relações semânticas, contribuindo para o fortalecimento da capacidade absortiva organizacional, entendida como a habilidade de reconhecer, adquirir, assimilar, transformar e explorar conhecimentos externos (Zahra & George, 2002). Esse processo favorece a geração de conhecimento estruturado e o avanço das práticas de análise tecnológica (O'Leary, 2013).

Entretanto, as ferramentas multilíngues de PLN enfrentam desafios significativos devido às complexidades da linguagem humana (Chen, 2024; Moraes et al., 2024) e ao domínio exercido por modelos voltados para as chamadas línguas de alto recurso, notadamente o inglês

(Kashyap, 2021). No campo da mineração textual de documentos de patente, soma-se a isso a estrutura técnica e complexa desses textos, que são multilíngues, semiestruturados e ricos em metadados (Shalaby & Zadrozny, 2019). Para o português, a limitação de recursos linguísticos impacta negativamente o desenvolvimento de modelos avançados de PLN (Almeida et al., 2024), restringindo os processos de aquisição e assimilação do conhecimento externo.

Estudos prévios indicam que abordagens puramente léxicas são insuficientes para capturar conceitos tecnológicos-chave em textos de patentes (Choi et al., 2013; Park, Ree, et al., 2013), uma vez que os campos não estruturados contêm termos de natureza funcional e atributiva, centrais ao processo inventivo (Kim, Choi, et al., 2018). Essas limitações foram confirmadas no Estudo 1, que evidenciou o impacto negativo da variação terminológica no desempenho da mineração textual, recomendando a incorporação de vocabulário específico do domínio.

Nesse contexto, o Estudo 2 investigou metodologias baseadas nas ferramentas da TRIZ, as quais oferecem uma estrutura sistemática e replicável para a organização do conhecimento inventivo. A TRIZ, fundamentada em parâmetros, propriedades, funções e contradições, apresenta elevada aderência à lógica descritiva dos documentos de patente, contribuindo para a transformação da informação técnica em conhecimento estruturado e acionável. Assim, sua integração à mineração textual favorece tanto a assimilação quanto a transformação do conhecimento, dimensões centrais da capacidade absortiva.

Por fim, no Estudo 3, descreve-se o desenvolvimento de uma ontologia semântica funcional baseada nos efeitos físicos da TRIZ, modelada segundo o esquema Sujeito–Ação–Objeto (Russo, 2011). Essa ontologia busca apoiar a recuperação automatizada de inteligência técnica, atuando como base semântica para aplicações em IA. A questão de pesquisa que orienta este estudo é: como recuperar inteligência técnica de patentes utilizando uma ontologia fundamentada nos efeitos físicos da TRIZ? O objetivo consiste em desenvolver um método de mineração de inteligência técnica, de natureza não supervisionada, que amplie a capacidade absortiva e o uso estratégico das patentes como ativos de conhecimento, em consonância com os pressupostos da KBV.

6.2 Revisão da Literatura

A seção de revisão da literatura está organizada em quatro subseções. A primeira identifica e analisa os desafios da mineração textual de patentes, compilando os principais

achados dos Estudos 1 e 2 desta tese. A segunda apresenta o fluxo de trabalho na mineração textual, descrevendo as etapas desde o pré-processamento até a avaliação dos resultados.

6.2.1 Mineração Textual de Patentes: Desafios e Perspectivas no Cenário Global e Brasileiro

Ao longo dos estudos, evidencia-se que as patentes constituem uma ampla fonte de conhecimento científico e prático, abrangendo praticamente todos os domínios do saber. As informações extraídas desses documentos técnicos contribuem para a construção do conhecimento e, simultaneamente, apoiam e orientam aspectos tecnológicos, de mercado, bem como ações estratégicas e decisórias. No Estudo 1 desta tese, esses aspectos foram categorizados em dimensões, as quais reforçam o papel das patentes como instrumento estratégico para a inovação e para a gestão.

De modo geral, em patentes de qualquer idioma, a complexidade linguística e estrutural constitui talvez o aspecto mais crítico. Esses documentos desempenham uma dupla função: (i) descrever tecnicamente a invenção e (ii) delimitar, por meio das reivindicações, o escopo da proteção legal solicitada. (Berdyugina & Cavallucci, 2020a). A linguagem técnica, jurídica e altamente especializada (Ali et al., 2025; Krestel et al., 2023), distinta da linguagem geral, dificulta que modelos de PLN padrão, geralmente treinados em textos genéricos, interpretem e processem o conteúdo com precisão (Ali et al., 2025; Souili, Cavallucci, & Rousselot, 2015b).

Os documentos de patente são compostos por seções fixas (título, resumo, reivindicações, descrição detalhada e figuras), cada uma com objetivos distintos e diferentes níveis de densidade informacional (Abbas et al., 2014; Khadilkar et al., 2019). A descrição da invenção guarda maior proximidade com a redação científica, mas apresenta frases complexas e recorrentes repetições de informações, a fim de delimitar de forma precisa o escopo da proteção (Berdyugina & Cavallucci, 2021). Já as reivindicações utilizam uma linguagem de caráter jurídico, mais restritiva e formal. Em ambas as seções e no resumo, observa-se o uso intensivo de terminologia específica de domínio, além de sinônimos, homônimos e termos polissêmicos, que podem assumir diferentes significados conforme o contexto. Esse fenômeno favorece ambiguidades semânticas e compromete a precisão das análises (Chen et al., 2022; Kim et al., 2019). Ademais, os documentos de patentes tendem a ser extensos e conter detalhes técnicos minuciosos, o que torna seu processamento computacionalmente custoso e desafiador (Ali et al., 2025).

Um segundo aspecto que impacta a eficiência da mineração de textos de patentes são as diferenças inerentes entre os idiomas, incluindo variações nas estruturas gramaticais, no vocabulário e nos sistemas de escrita, que comprometem a precisão na extração e no

processamento de termos e dificultam a adaptação e a generalização dos métodos desenvolvidos (Krishna, 2023; Liwei, 2022). Nesse contexto, destacam-se as chamadas linguagens de alto recurso que dominam o desenvolvimento de modelos de PLN, enquanto as linguagens de baixo recurso (entre as quais se inclui o português) permanecem desassistidas (Kashyap, 2021).

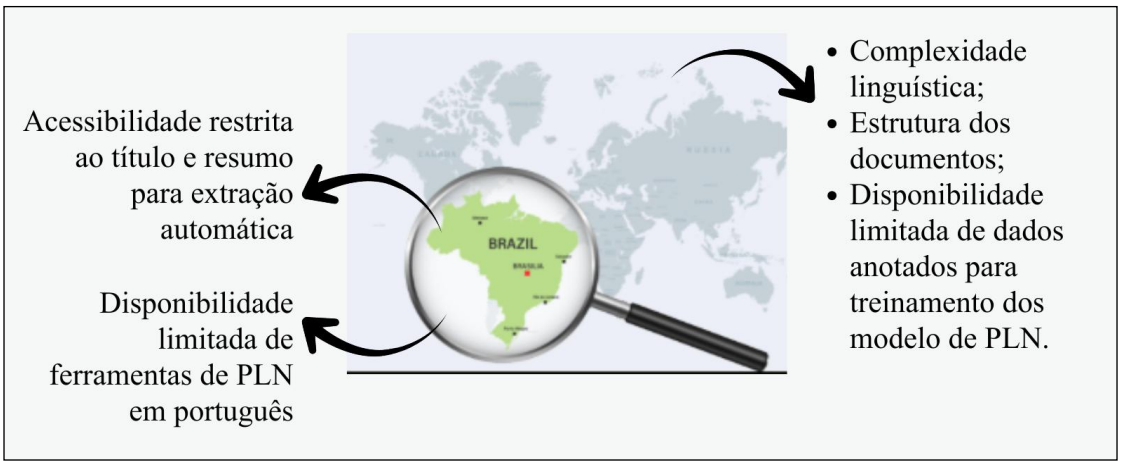
A língua portuguesa, classificada como de baixo recurso, apresenta disponibilidade limitada de ferramentas de PLN. A adaptação de sistemas originalmente desenvolvidos para línguas de alto recurso geralmente não produz resultados satisfatórios, em razão das ambiguidades semânticas, sintáticas e lexicais do português (Moraes et al., 2024), além das significativas variações regionais — ao contrário do inglês, idioma relativamente mais uniforme (Azevedo, 2005). Para avaliar essa questão, na seção de apresentação dos resultados deste estudo é descrito o *setup* experimental que comprova a perda de precisão dos modelos treinados em inglês quando aplicados a textos em português.

Os dados de treinamento para modelos de PLN, que constituem a base para sistemas de IA capazes de reconhecer padrões linguísticos, são obtidos em bases de conhecimento como bases patentárias, classificadores de patentes ou em outras fontes não patentárias, compondo os chamados dados brutos (Huang & Xie, 2022; Liu et al., 2023). No entanto, em algumas jurisdições — entre elas o Brasil — a acessibilidade restrita aos dados textuais de patentes representa uma barreira para a análise automática. Frequentemente, apenas o título e o resumo estão disponíveis em formato adequado para extração automatizada, enquanto a descrição detalhada da invenção e as reivindicações permanecem menos acessíveis, muitas vezes disponibilizadas apenas em arquivos digitalizados.

Quanto aos dados anotados (textos previamente marcados por humanos com categorias ou rótulos), além de demandarem alto custo e tempo considerável de especialistas no domínio, sua disponibilidade permanece limitada (Guarino et al., 2021; Xu et al., 2019). Por esse motivo, métodos não supervisionados ou semissupervisionados têm prevalecido nas estratégias de mineração de textos de patentes em âmbito global, conforme apontado no Estudo 1 desta tese. A Figura 19 apresenta os principais desafios inerentes à mineração textual de patentes, tanto no cenário mundial quanto no contexto brasileiro.

Figura 19

Desafios globais e brasileiros na mineração textual de patentes



Nota. Dados da pesquisa (2025).

Este cenário desafiador impacta diretamente os estudos de mineração textual de patentes, sobretudo em idiomas de baixo recurso, como a língua portuguesa. Ainda assim, constitui uma área de investigação relevante, considerando que o domínio das patentes reúne um volume expressivo de dados — milhões de documentos disponíveis. Desse modo, avançar em metodologias capazes de lidar com tais especificidades torna-se essencial para ampliar a precisão e a eficiência das análises automatizadas.

6.2.2 Fluxo de Trabalho na Mineração Textual de Patentes

O fluxo de trabalho na mineração textual de patentes segue uma sequência de etapas, que compreende: a obtenção dos documentos de patente; o pré-processamento dos textos; a representação dos textos; a extração de padrões e informações; e a avaliação dos resultados.

6.2.2.1 Obtenção dos documentos de patente

Para a obtenção dos documentos de patente, realizam-se pesquisas em bases patentárias. Além das bases nacionais, que reúnem documentos depositados por inventores locais ou por aqueles que buscam proteção no país, existem bases regionais, como a EAPO (*Eurasian Patent Organization*), que abrange países da Eurásia; a ARIPO (*African Regional Intellectual Property Organization*), que cobre alguns países da África Anglófona[12]; e a OAPI (*Organisation Africaine de la Propriété Intellectuelle*), que integra países da África Francófona[13].

[12] Conjunto de países africanos cuja língua oficial é o inglês. Exemplos: Nigéria, Gana, Quênia, Uganda, África do Sul, Tanzânia, Zâmbia, Zimbábue.

[13] Conjunto de países africanos cuja língua oficial é o francês**.** Exemplos: Senegal, Costa do Marfim, Mali, Níger, Burkina Faso, Chade, Benin, República Democrática do Congo, Camarões (parcialmente francófono)**.**

Adicionalmente, destacam-se as bases de abrangência internacional, que reúnem milhões de documentos de diferentes jurisdições em uma única interface. Entre elas, podem ser mencionadas: a *Espacenet*, mantida pelo Escritório Europeu de Patentes, com um repositório de mais de 151 milhões de documentos (European Patent Office, 2025a); a *Patentscope*, da Organização Mundial da Propriedade Intelectual (*World Intellectual Property Organization* - WIPO), com cerca de 123 milhões de patentes (World Intellectual Property Organization, 2025b); além da *Google Patents* e da plataforma *Lens.org*, todas de acesso público.

Entre as bases de acesso comercial, destacam-se a *Derwent Innovation*, a *Orbit Intelligence*, a *PatBase,* a *LexisNexis TotalPatent One* e a *Innography*.

Nesta etapa de coleta dos documentos, são definidos os campos textuais a serem analisados, que podem incluir título, resumo, descrição detalhada, reivindicações e classificações de patente, como a IPC e a Classificação Cooperativa de Patentes (CPC). Os textos brutos coletados passam, então, por um processo de estruturação em formato adequado, na etapa de pré-processamento.

6.2.2.2 Pré-processamentos dos textos brutos dos documentos de patente

Na etapa de pré-processamento, o texto bruto da patente é preparado para análise computacional. Essa etapa é fundamental para reduzir o tamanho do conjunto de dados e transformá-los em informações estruturadas, facilitando o processo de mineração (Chan et al., 2021; Rüdiger et al., 2017).

Entre as abordagens de pré-processamento convencionalmente aplicadas à mineração de textos genéricos e de patentes, destaca-se a **tokenização**. Esse procedimento é essencial para estruturar a linguagem técnica densa presente em documentos patentários, tornando-a passível de análise automática. Nos Estudos 1 e 2 desta tese, verificou-se que a eficácia dos modelos linguísticos aplicados à mineração textual de patentes depende diretamente do alinhamento entre o método de tokenização e as características linguísticas do domínio, o que reforça a relevância dessa abordagem (Althammer et al., 2021; Chan et al., 2021; Vereschak & Korobkin, 2019).

A tokenização envolve etapas como a conversão do texto para letras minúsculas (Antons et al., 2016), a divisão em sentenças e o fracionamento de cada sentença em unidades ortográficas denominadas *tokens* (Antons et al., 2020; Chan et al., 2021; Chiarello et al., 2018). Entre as ferramentas mais utilizadas destacam-se a biblioteca de programação em Python NLTK (*Natural Language Toolkit*), que disponibiliza módulos específicos para tokenização (Chan et al., 2021); a ferramenta *SpaCy*, voltada para a extração de entidades nomeadas, análise de

dependências sintáticas e vetorização semântica (Russo et al., 2018); e os modelos de linguagem baseados em transformadores, como o BERT e o SciBERT (Althammer et al., 2021). Neste último caso, Althammer et al. (2021) demonstram que a tokenização SciBERT apresenta desempenho superior à do BERT padrão para textos de patentes, pois gera frases codificadas mais curtas e reduz a taxa de divisão.

A **remoçãode palavras irrelevantes (*stopwords*)** contribui para aumentar a eficiência e a precisão da análise textual. Esse processo consiste em identificar e remover termos que, embora frequentes, possuem baixo valor semântico, como preposições, conjunções e pronomes (Korobkin, Fomenkov, & Kravets, 2018; Vereschak & Korobkin, 2019; Yoon et al., 2022). No contexto de patentes, certas palavras também se repetem de forma recorrente, mas não agregam significado único (Berdyugina & Cavallucci, 2022b; Khode, 2019), como "invenção" ou o bigrama "caracterizado por".

Para a verificação de palavras irrelevantes, a literatura aponta estratégias como a ponderação pela frequência inversa do documento (TF-IDF) (Wang & Liu, 2024) ou a definição de parâmetros de "amostragem reduzida", baseados na frequência de termos, aplicados em modelos de *machine learning* como o Word2Vec (Sarica et al., 2020). Contudo, a seleção de palavras irrelevantes apenas pela frequência de ocorrência envolve o risco de excluir termos que, embora recorrentes, sejam portadores de significado relevante (Rose et al., 2010).

A **lematização** consiste na redução de palavras flexionadas ou derivadas à sua forma básica (*lema*), preservando uma forma gramaticalmente válida (Berdyugina & Cavallucci, 2023; Chan et al., 2021; Chiarello et al., 2018; Kim, Choi, et al., 2018). Por exemplo, a forma básica de "corredor" e "corrida" é "correr". Entre as ferramentas comumente utilizadas, destacam-se a biblioteca NLTK e o *WordNet Lemmatizer*, que se apoia no banco lexical *WordNet* (Chan et al., 2021).

A **stemização** reduz palavras flexionadas ou derivadas ao seu tronco (*stem*), base ou raiz, que nem sempre corresponde à raiz morfológica, podendo gerar formas não gramaticalmente corretas (Berdyugina & Cavallucci, 2022b; Rossi et al., 2019). Por exemplo, "corre", "correndo" e "correu" são reduzidos ao tronco "corr".

A **derivação** busca identificar a "base da palavra" a partir da remoção de afixos (prefixos e sufixos). Diferentemente da lematização, a base resultante nem sempre corresponde à raiz morfológica ou a uma palavra real, mas sim a um fragmento da forma original (Berdyugina & Cavallucci, 2022b; Chan et al., 2021). Por exemplo, "jogar" é considerado a forma base de variações como "joga", "brinca" e "jogo".

Em síntese, o pré-processamento de textos de patentes é responsável por limpar e

estruturar os dados textuais, permitindo que os algoritmos de mineração operem sobre informações significativas e de maior qualidade.

6.2.2.3 Representação dos textos

Esta etapa consiste em converter os dados não estruturados das patentes em um formato estruturado, adequado para processamento por algoritmos de mineração de texto. As formas de representação podem ser agrupadas em três grandes categorias: representações baseadas em vetores, representações por tópicos e representações híbridas.

As **representações baseadas em vetores** descrevem palavras e frases como vetores em um espaço multidimensional, posicionando termos semanticamente próximos em regiões vizinhas desse espaço. A proximidade e a orientação dos vetores refletem a relevância semântica ou o grau de similaridade entre os termos (Sarica et al., 2020; Teng et al., 2024; Wang & Liu, 2024). Entre as principais abordagens de PLN utilizadas nesse domínio — e discutidas nos Estudos 1 e 2 desta tese — destacam-se:

(a) Saco de palavras (*Bag-of-Words* - BoW) - representa cada documento como um vetor de frequências de palavras, desconsiderando a ordem em que aparecem (Miric et al., 2023). Apesar de simples, o método não captura contexto ou significado, tratando como distintas palavras semanticamente semelhantes escritas de forma diferente e como idêntica a mesma palavra em diferentes sentidos (Chan et al., 2021; Miric et al., 2023);

(b) TF-IDF - pondera o peso das palavras com base em sua frequência no documento e no corpus. Quanto mais rara no conjunto geral e mais frequente em um documento específico, maior a pontuação, indicando termos potencialmente relevantes como palavras-chave (Liu et al., 2021; Wu, 2019);

(c) Modelos de incorporação de palavras (*Word2Vec, GloVe, FastText*) - reduzem a dimensionalidade da representação textual, projetando palavras em vetores densos que preservam relações semânticas. Dessa forma, palavras usadas em contextos semelhantes tendem a se posicionar próximas no espaço vetorial (Jeon et al., 2024; Li et al., 2024; Sarica et al., 2020);

(d) *Embeddings* contextuais - modelos de linguagem como *BERT, SciBERT* e *PatentBERT* geram representações dependentes do contexto, capturando de forma mais precisa nuances semânticas. Enquanto o *BERT* é pré-treinado em linguagem de uso geral, como o *Wikitext* (Jeon et al., 2024), o *SciBERT* é especializado em literatura científica (Althammer et al., 2021; Jeon et al., 2024)

e o *PatentBERT* é um modelo de linguagem especializado para patentes (Althammer et al., 2021).

A **representação por tópicos** transforma coleções de textos em estruturas mais compactas, capazes de capturar os temas latentes (ou tópicos) presentes nos documentos. Diferentemente das abordagens vetoriais, que representam diretamente palavras ou sentenças, a representação por tópicos busca identificar a distribuição de tópicos em cada documento e a distribuição de palavras em cada tópico, revelando padrões abstratos e estruturas semânticas ocultas em grandes coleções textuais (Trappey et al., 2024).

Entre as metodologias mais utilizadas destacam-se:

(a) Alocação Latente de Dirichlet (LDA) - modelo generativo probabilístico amplamente aplicado para classificar textos em tópicos específicos (Joshi et al., 2022; Wei et al., 2023). É frequentemente empregado para identificar automaticamente tópicos tecnológicos a partir de textos de patentes (Tian et al., 2022);

(b) Análise Semântica Latente (*Latent Semantic Analysis* – LSA) - técnica que converte o conteúdo semântico em representações vetoriais, extraindo o significado contextual das palavras por meio de cálculos estatísticos aplicados a grandes *corpora* textuais (Chen et al., 2020; Trappey et al., 2023; Zhang et al., 2020);

(c) Análise Semântica Latente Probabilística (*Probabilistic Latent Semantic* Analysis - PLsA) - abordagem que interpreta documentos como distribuições probabilísticas de tópicos, cada um representado por distribuições de palavras (Berdyugina & Cavallucci, 2020b). Apesar de reconhecida como técnica de modelagem de tópicos, os métodos examinados nos Estudos 1 e 2 não identificaram trabalhos que a utilizassem de forma ativa.

As **representações híbridas ou enriquecidas** constituem abordagens que combinam distintas formas de representação de texto, como métodos baseados em vetores, tópicos, embeddings e informações externas (ontologias ou metadados), com o objetivo de capturar simultaneamente características estatísticas e significados semânticos mais profundos. Nos estudos analisados, observa-se a aplicação de estratégias que integram:

(a) Técnicas estatísticas, como o TF-IDF, em conjunto com *embeddings* (*Word2Vec, FastText*), de modo a enriquecer a análise ao considerar tanto a frequência de termos quanto suas relações semânticas (Chen et al., 2020);

(b) Modelos de tópicos associados a *embeddings*, ampliando a capacidade de

extração de padrões (Jiang et al., 2025; Wei et al., 2023);

(c) Ontologias, taxonomias ou redes semânticas, incluindo classificações de patentes, como IPC e CPC, para capturar relações hierárquicas e conceituais entre termos (Kitamura et al., 2024; Sarica et al., 2020; Taduri et al., 2019; Trappey et al., 2023).

Uma vez que os textos são convertidos em representações estruturadas, seja em forma de vetores, tópicos ou modelos híbridos, torna-se possível aplicar algoritmos capazes de identificar relações relevantes, regularidades e estruturas ocultas nos dados. Essa etapa marca a passagem da codificação dos conteúdos textuais para a extração de padrões e informações, onde o foco deixa de ser a organização do texto e passa a ser a descoberta de conhecimento significativo.

6.2.2.4 Extração de padrões e informações

A extração de padrões e de informações em mineração textual é a etapa em que os dados de texto, representados em um formato estruturado (vetores, tópicos ou híbridos), passam a ser analisados por algoritmos para revelar conhecimento oculto, tendências e relações significativas. Essa análise pode ser realizada por diferentes métodos e técnicas, muitas vezes utilizadas de forma complementar.

A **extração de termos e entidades** consiste em identificar automaticamente, dentro de grandes volumes de texto, as palavras ou expressões mais relevantes (termos ou palavras-chave) que descrevem o conteúdo técnico da patente, bem como os elementos com significado específico, processo conhecido como reconhecimento de entidades nomeadas (*Named Entity Recognition* – NER). Para isso, pode-se recorrer a marcadores morfossintáticos (*Part-of-Speech tagging* – PoS), que identificam categorias gramaticais (substantivos, verbos, adjetivos, advérbios etc.), ou ainda à detecção de *n*-gramas (sequências de palavras) recorrentes no texto (Sarica et al., 2020).

Exemplo 1 – Trecho fictício de uma patente de invenção e respectivos dados bibliográficos:

A presente invenção refere-se a um dispositivo de resfriamento termoelétrico para uso em veículos elétricos. O sistema utiliza nanotubos de carbono para melhorar a condutividade térmica. O inventor João Silva, associado à Universidade X, depositou a patente em 12 de agosto de 2023, classificada na CPC H01L 35/32.

A partir do trecho acima, podem ser extraídos os seguintes elementos:

Categoria	Elementos extraídos
Termos técnicos	dispositivo de resfriamento termoelétrico; veículos elétricos; nanotubos de carbono; condutividade térmica
Entidades nomeadas	Inventor: João Silva; Instituição: Universidade X; Data: 12 de agosto de 2023; Classificação: CPC H01L 35/32

A **descoberta de padrões léxicos e semânticos** vai além da identificação de termos isolados, buscando reconhecer regularidades, associações e relações de significado no discurso técnico. Para essa finalidade, diversas ferramentas linguísticas podem ser utilizadas. O NLTK **(***Natural Language Toolkit***)**, por exemplo, realiza a tokenização, segmentando o texto em palavras e frases para análise lexical (Chan et al., 2021; Giordano et al., 2024). O *SpaCy* permite a extração de entidades nomeadas (NER), análise de dependências sintáticas e vetorização semântica (Russo et al., 2018; Zhou et al., 2024). Já o *Stanford CoreNLP* possibilita a extração de estruturas SAO (Li et al., 2023) e a identificação de relações palavra a palavra em uma frase (Vereschak & Korobkin, 2019; Yoon et al., 2022).

Exemplo 2 - Utilizando o mesmo trecho de patente do Exemplo 1, podem ser extraídos:

Categoria	Elementos extraídos
Padrão léxico	Bigramas: resfriamento termoelétrico; veículos elétricos; nanotubos de carbono; condutividade térmica
Padrão semântico	Relação de função e propósito: [dispositivo de resfriamento termoelétrico → uso → veículos elétricos] Relação de tecnologia empregada: [sistema → utiliza → nanotubos de carbono] Relação de causa e efeito: [nanotubos de carbono → melhoram → condutividade térmica]

O **agrupamento (ou clusterização)** permite organizar patentes semelhantes em grupos (*clusters*) de acordo com o grau de similaridade entre elas (Berdyugina & Cavallucci, 2022b). Entre os algoritmos mais frequentemente citados nos estudos de mineração de textos de patentes, destacam-se o *K-means* (Kim & Yoon, 2022; Trappey et al., 2024) e os métodos de modelagem de tópicos, como o LDA (Berdyugina & Cavallucci, 2023; Wei et al., 2023) e o *BERTopic* (Jiang et al., 2025), que geram agrupamentos temáticos ao extrair tópicos latentes das patentes. Essa organização facilita a identificação de subáreas tecnológicas, a detecção de tendências emergentes e a visualização de paisagens tecnológicas.

Além disso, Maskittou et al. (2022) utilizam a técnica de fatoração matricial por Decomposição de Valor Singular (*Singular Value Decomposition* – SVD) como recurso de redução de dimensionalidade e modelagem de tópicos em textos de patentes, permitindo revelar estruturas temáticas subjacentes.

Exemplo 3 - Amostra de patentes relacionadas a tecnologias de armazenamento de energia e sensoriamento

P1: Método para aumentar a eficiência de baterias de íon-lítio com novos eletrólitos sólidos.
P2: Dispositivo de recarga rápida para baterias de veículos elétricos.
P3: Sensor óptico para monitoramento ambiental de poluentes.
P4: Sensor de pressão para aplicações médicas em monitoramento cardíaco.
P5: Bateria flexível para dispositivos vestíveis baseada em polímeros condutivos.
P6: Sistema híbrido de sensor químico e óptico para detecção de gases tóxicos.

A partir da amostra de patentes acima, podem ser gerados pelo menos dois agrupamentos:

Agrupamento	Patentes selecionadas
Patentes sobre baterias	P1 (eficiência de baterias) P2 (recarga rápida) P5 (bateria flexível)
Patentes sobre sensores	P3 (sensor óptico ambiental) P4 (sensor de pressão médico) P6 (sensor químico + óptico)

A **análise de similaridade** é empregada para mensurar o grau de proximidade semântica ou lexical entre documentos ou termos. Essa técnica viabiliza o agrupamento de elementos textuais, revelando subáreas tecnológicas e permitindo a identificação de sinônimos e variações terminológicas. Além disso, contribui para a detecção de relações entre conceitos técnicos em redes semânticas e para a extração de tópicos latentes por meio de métodos de modelagem de tópicos (Helmers et al., 2019; Li, Wang, et al., 2023; Teng et al., 2024; Yoon et al., 2022).

Exemplo 4 - Considerando o mesmo conjunto de patentes do Exemplo 3, pode-se extrair como padrão de similaridade semântica:

Agrupamento	Patentes selecionadas
Baterias e cargas	P1 P2 P5
Sensores	P3 (sensor óptico ambiental) P4 (sensor de pressão médico) P6 (sensor químico + óptico)

A **modelagem de tópicos** é um modelo estatístico que pode operar como um método de aprendizado não supervisionado que agrupa estruturas semânticas implícitas em um conjunto de texto (Li et al., 2024; Ma et al., 2021; Trappey, Lin, et al., 2024), permitindo descobrir tópicos ocultos em um corpo de texto sem necessariamente demandar uma parcela de dados da base anotados para que o modelo possa ser treinado (Zhang et al., 2018). No contexto da mineração textual, diferentes modelos estatísticos e probabilísticos têm sido aplicados para extrair padrões semânticos latentes. O LDA, por exemplo, assume que os documentos são composições de

múltiplos tópicos latentes, cada um caracterizado por uma distribuição de palavras-chave, permitindo identificar estruturas temáticas em grandes *corpora* (Trappey et al., 2024). Já a Indexação Semântica Latente Probabilística (PLSi) modela a coocorrência entre termos e documentos por meio de distribuições probabilísticas mediadas por variáveis latentes, favorecendo uma interpretação mais robusta das associações semânticas (Ma et al., 2021). Por sua vez, o método de Explicação da Correlação (*Correlation Explanation* – CorEx) busca identificar agrupamentos de variáveis maximamente informativos, revelando relações não triviais entre termos e documentos, o que amplia a capacidade de descoberta de conhecimento em bases textuais complexas (Trappey et al., 2024),

Exemplo 5 – Exemplo fictício de um resumo de patente

Um sistema de purificação de água inclui um módulo de filtração por membrana, um tanque de retenção de sólidos e sensores de pressão para monitoramento em tempo real. O sistema utiliza bombas de alta eficiência e válvulas automatizadas para controlar o fluxo. A invenção visa aumentar a eficiência energética e reduzir custos operacionais em processos industriais.

A partir do trecho acima, podem ser extraídos tópicos, por meio de técnicas de modelagem de tópicos utilizando LDA ou PLSi:

Tópico	Palavras-chave extraídas
Eficiência e custo operacional	eficiência, energia, reduzir, custo, processo, industrial
Filtração e tratamento de água	água, filtração, membrana, purificação, tanque, sólido
Automação e controle de fluxo	válvula, bomba, fluxo, pressão, sensor, monitoramento

Os modelos de IA generativa têm sido aplicados à mineração de texto para executar tarefas complexas que vão além da mera extração de padrões lexicais, introduzindo capacidades interpretativas e preditivas. Entre os exemplos, destacam-se o uso de GANs para a sumarização de documentos de patentes, capturando suas principais informações técnicas (Kim & Yoon, 2022), e o mapeamento topográfico generativo (*Generative Topographic Map* - GTM) para a identificação de lacunas tecnológicas ou áreas ainda inexploradas (Liu et al., 2023).

Mais recentemente, os LLMs têm ganhado relevância na mineração textual de patentes por superarem limitações dos métodos tradicionais de análise. Seus recursos avançados permitem lidar com a complexidade estrutural e vocabular dos textos, além de processar grandes volumes de documentos, sendo aplicados desde o pré-processamento até a extração de conceitos inventivos complexos e a realização de buscas semânticas (Jiang et al., 2025; Trapp & Warschat, 2025). Considerando a recorrente escassez de dados anotados no domínio das patentes, os LLMs (como o GPT-4) apresentam-se como uma alternativa promissora para

superar essa limitação e viabilizar a extração de conhecimento técnico de forma mais eficiente (Blume et al., 2024; Trapp & Warschat, 2025).

A extração de padrões e informações em mineração textual pode ser realizada por meio de diferentes métodos e técnicas, muitas vezes aplicados de forma combinada. Entretanto, as relações hierárquicas e semânticas entre conceitos tendem a se perder nesse processo, o que compromete a profundidade da análise. Nos Estudos 1 e 2 desta tese, voltados à avaliação de ferramentas de mineração textual de patentes, observou-se que a integração de ontologias possibilita resultados mais robustos e interpretativos, viabilizando a descoberta de relações técnicas complexas, como os vínculos entre problemas e soluções ou entre funções e tecnologias, o que amplia significativamente o potencial de exploração do conhecimento contido nos documentos de patente.

6.2.2.5 Avaliação dos resultados

Para avaliar o desempenho dos métodos de mineração de texto, particularmente no contexto da análise de patentes, os pesquisadores empregam diferentes métricas quantitativas, cuja escolha depende da tarefa específica, como classificação, agrupamento ou medição de similaridade (Miric et al., 2023; Wang & Liu, 2024). De forma geral, essas métricas podem ser agrupadas em três categorias, conforme a finalidade.

A primeira categoria corresponde às **métricas para qualidade da geração de texto**, aplicadas em tarefas como resumo automático e tradução de textos. Entre as mais utilizadas estão:

(a) ROUGE (*Recall-Oriented Understudy for Gisting Evaluation*) - avalia a qualidade de sistemas de geração de linguagem natural, como resumos automáticos e traduções, comparando os textos gerados com referências humanas. Essa métrica se concentra em dois aspectos principais: o *recall*, que mede até que ponto o conteúdo do texto de referência foi recuperado, e a precisão, que avalia a relevância das palavras incluídas no texto gerado (Kim & Yoon, 2022; Trappey, Trappey, Wu, et al., 2020);

(b) BLEU (*Bilingual Evaluation Understudy*) - amplamente empregada na avaliação de traduções automáticas e resumos, baseia-se na semelhança lexical entre o texto gerado e um ou mais textos de referência. Essa métrica considera a ordem das palavras e a sobreposição de *n*-gramas, permitindo medir a proximidade linguística entre as produções automáticas e as versões humanas (Kim & Yoon, 2022).

As **métricas para recuperação e classificação de informações** são utilizadas para avaliar o desempenho de métodos aplicados à categorização de documentos em classes predefinidas, geralmente por meio da comparação entre as previsões do modelo e um conjunto de dados rotulado manualmente (Ali et al., 2025; Guarino et al., 2022; Kang et al., 2018; Lee & Bai, 2025; Miric et al., 2023; Puccetti et al., 2021; Zhang et al., 2024). No caso específico de tarefas de recuperação da informação, as métricas medem a capacidade de um modelo em ranquear documentos relevantes acima dos irrelevantes (Lee & Bai, 2025), sendo fundamentais para comparar diferentes abordagens e compreender sua eficácia em aplicações específicas (Miric et al., 2023).

As métricas fundamentais compreendem:

(a) Acurácia – mede a proporção de documentos corretamente classificados (tanto verdadeiros positivos (VP) quanto verdadeiros negativos (VN)) em relação ao total de documentos avaliados (Ali et al., 2025). Representa uma visão geral do desempenho do modelo e é definida pela Equação 1:

$$\text{Acurácia} = \frac{\text{número de registros VP} + \text{VN}}{\text{total de registros avaliados}} \quad (1)$$

(b) Precisão – mede a proporção de documentos verdadeiros positivos em relação ao total de documentos classificados como positivos pelo modelo (VP) e falsos positivos (FP)). Essa métrica indica a qualidade dos resultados recuperados, ou seja, a fração dos documentos que o modelo identificou como relevantes e que de fato o são (Ali et al., 2025; Hu et al., 2018). É definida pela Equação 2:

$$\text{Precisão} = \frac{\text{número de registros VP}}{\text{número de registros VP} + \text{FP}} \quad (2)$$

(c) *Recall* - mede a proporção de resultados VP em relação ao total de documentos que deveriam ter sido identificados como positivos (VP + FN). Essa métrica expressa a capacidade do modelo de recuperar todos os documentos relevantes. Maximizar o *recall* é particularmente importante em tarefas de mineração textual e recuperação de informação, pois garante que os resultados da pesquisa capturem o maior número possível de documentos relevantes, reduzindo o risco de perda de informações importantes (Ali et al., 2025; Zhang et al. 2018). O *recall* é definido pela Equação 3:

$$Recall = \frac{\text{número de registro VP}}{\text{número de registros VP} + \text{FN}} \quad (3)$$

(d) F1-*score* – fornece uma métrica que captura o equilíbrio entre a precisão de um modelo (proporção de itens selecionados que são relevantes) e seu *recall* (proporção de itens relevantes que são recuperados). Caracteriza-se por ser a média harmônica entre precisão e *recall*, sendo especialmente útil em contextos de mineração textual com desbalanceamento de classes (como quando poucos documentos relevantes estão entre muitos irrelevantes). Nesses casos, o F1-*score* fornece uma medida única e equilibrada entre a capacidade do modelo de recuperar documentos relevantes e de evitar falsos positivos (Zhang et al., 2018). É definido pela Equação 4:

$$F1 - score = 2\,\frac{\text{precisão x } recall}{\text{precisão} + recall} \tag{4}$$

(e) *Exact Match* (EM) ou Acurácia de Subconjunto - em contextos de aprendizado de máquina, mede a proporção de previsões que correspondem exatamente ao rótulo (Lee & Bai, 2025; Zheng et al., 2024). É definido pela Equação 5:

$$Exact\ Match = \frac{\text{número de respostas corretas}}{\text{total de respostas}} \tag{5}$$

Para tarefas especializadas, podem ser necessárias métricas adaptadas ou desenvolvidas especificamente para refletir com maior precisão o desempenho e atender a objetivos particulares de pesquisa. Entre elas, destacam-se:

(a) Métrica para extração de contradições – avalia a capacidade de sistemas de mineração textual em identificar contradições técnicas em documentos de patente (Guarino, Samet, Nafi, et al., 2020);

(b) Métricas de convergência tecnológica – são utilizadas para mensurar a combinação inicial de duas classificações internacionais de patentes, permitindo analisar tendências de convergência entre diferentes áreas tecnológicas (Wang et al., 2024);

(c) Métrica de coerência de tópicos – também chamada de pontuação de consistência, é utilizada na modelagem de tópicos (como LDA) para avaliar a qualidade e a coerência dos tópicos gerados (Nkolongo et al., 2024);

(d) Informação Mútua Pontual Normalizada (*Normalized Pointwise Mutual Information* - NPMI) – aplicada na avaliação de modelos de tópicos, mensura a qualidade semântica da associação entre termos. Valores mais altos indicam maior coerência e melhor desempenho (Jiang et al., 2025);

(e) Pontuação de homogeneidade – empregada na avaliação de algoritmos de

agrupamento, verifica se cada cluster resultante contém apenas documentos pertencentes a uma mesma classe predefinida, sendo particularmente útil na classificação de patentes (Wang & Liu, 2024).

Também são previstas **métricas de eficiência**, que avaliam a aplicação prática dos métodos de mineração de texto em termos de desempenho computacional:

(a) Tempo de treinamento - corresponde ao tempo necessário para criar um modelo estatístico a partir de um conjunto de dados de treinamento (Puccetti et al., 2021; Teng et al., 2024);

(b) Tempo de extração - refere-se ao tempo requerido para que um sistema identifique e extraia as entidades-alvo de um conjunto de documentos de patente (Chiarello et al., 2018; Puccetti et al., 2021);

(c) Custo computacional – diz respeito à sobrecarga de processamento envolvida na execução do método, fator que impacta diretamente a escalabilidade da aplicação (Ali et al., 2025).

Por fim, podem ser aplicadas **métricas subjetivas de avaliação**, em complemento às métricas convencionais, que permitem a especialistas humanos avaliar a utilidade prática dos sistemas de mineração de texto, geralmente por meio de escalas de percepção ou questionários estruturados (Kitamura et al., 2024; Teng et al., 2024; Wang & Liu, 2024).

Após a apresentação do fluxo geral de mineração textual de patentes, será detalhada a metodologia adotada no presente estudo, descrevendo de forma sistemática cada etapa do processo. São explicitados os procedimentos de coleta e preparação dos dados, a modelagem da ontologia, as técnicas de processamento de linguagem natural empregadas e os critérios utilizados para a extração e análise das soluções técnicas. Essa descrição visa oferecer transparência metodológica, permitindo a reprodutibilidade do estudo e fundamentando a análise dos resultados obtidos.

6.3 Procedimentos Metodológicos

O presente estudo adota uma abordagem qualitativa, de natureza exploratória (Patton, 2015), voltada à documentação sistemática dos resultados e à coleta de informações descritivas sobre o desempenho de um método de mineração textual aplicado a documentos de patente. Para a condução da análise, utiliza-se a ontologia desenvolvida no Estudo 3 como base semântica estruturada, destinada à organização de conceitos tecnológicos e à identificação de

suas inter-relações. Essa ontologia não apenas explicita vínculos conceituais e funcionais, como também incorpora regras sintáticas formais que possibilitam a extração de soluções técnicas genéricas ou conceituais diretamente a partir dos textos analisados.

A primeira subseção detalha o fluxo de trabalho experimental planejado, fundamentado nas etapas convencionais de mineração textual e associado às configurações experimentais, tais como as estabelecidas em Chen et al. (2020), Chiarello et al. (2018) e Jiang et al. (2025), contemplando quatro alternativas de técnicas alinhadas à evolução metodológica das abordagens de mineração textual utilizando TRIZ, conforme observado no Estudo 2 desta tese. A segunda subseção descreve os procedimentos de obtenção dos documentos de patente no INPI, enquanto a terceira subseção expõe o processo de avaliação dos resultados do método de mineração textual aplicado a documentos.

6.3.1 Fluxo de Trabalho Experimental e Justificativa

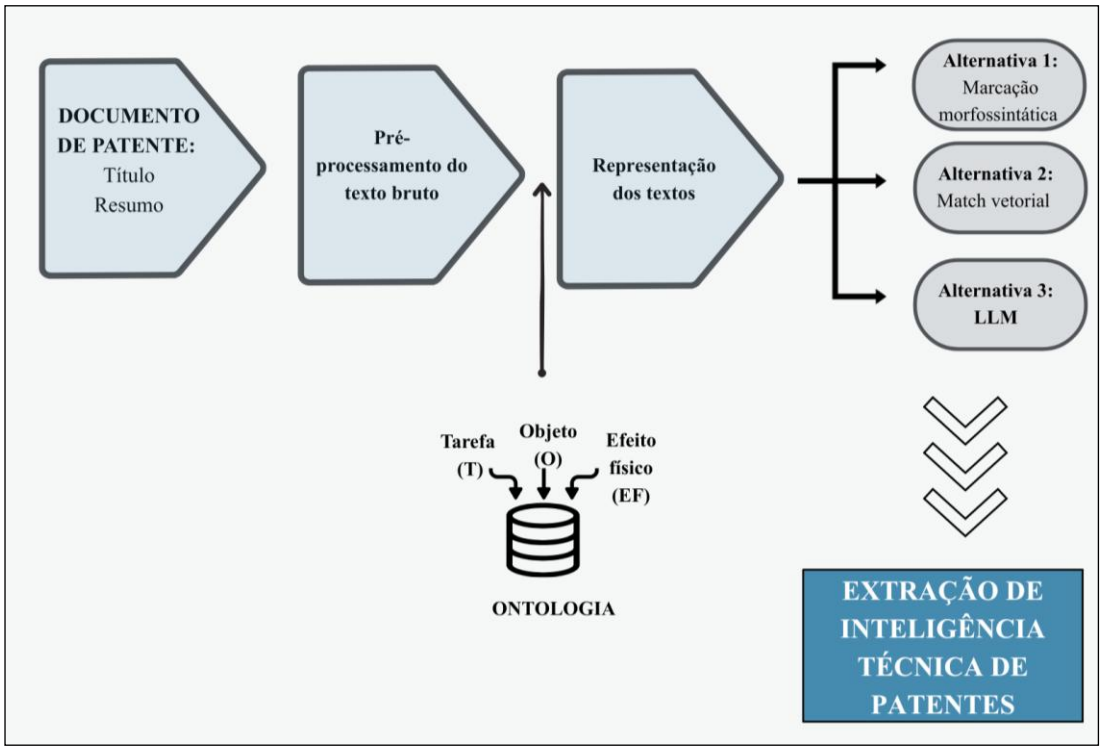
Com base no conhecimento acumulado nos Estudos 1 e 2 desta tese, foi estruturado um fluxo de trabalho experimental para a definição do modelo de mineração textual, o qual passou por ajustes ao longo dos experimentos em função dos resultados parciais obtidos. Esses ajustes serão descritos em detalhe na seção destinada à apresentação dos resultados.

O fluxo foi concebido para contemplar três macroetapas (pré-processamento, representação dos textos e extração da inteligência técnica), organizadas de forma a apoiar a mineração textual de patentes, tendo como base semântica a ontologia detalhada no Estudo 3.

Além disso, o fluxo experimental foi planejado de maneira progressiva: inicia-se com técnicas clássicas de PLN e avança para arquiteturas mais sofisticadas, baseadas em LLMs. Essa progressão justifica-se pela intenção de estabelecer uma linha comparativa entre métodos tradicionais e soluções de última geração, possibilitando a identificação de seus respectivos limites e potenciais no desafio de mineração semântica de patentes em relação à ontologia TRIZ, em consonância com a evolução metodológica observada no Estudo 2 desta tese. A Figura 20 ilustra o fluxo de trabalho planejado.

Figura 20

Fluxograma das etapas de desenvolvimento do método de mineração textual aplicado a patentes com suporte da ontologia



Nota. Elaborado pela Autora (2025).

Preliminarmente, foram definidas três alternativas metodológicas. Inicialmente, como etapa exploratória e com o objetivo de fornecer uma visão preliminar sobre a estrutura linguística das patentes, realizou-se, na **primeira alternativa experimental**, a anotação morfossintática (*PoS Tagging*), voltada à identificação das funções gramaticais de cada termo do corpus analisado. Essa etapa apoia-se em estudos anteriores que argumentam ser tal abordagem particularmente útil em textos técnicos, nos quais substantivos e verbos concentram a maior parte da informação conceitual (Chiarello et al., 2018; Guarino et al., 2022; Jiang et al., 2025).

Como **segunda alternativa experimental**, os termos do corpus analisado foram vetorizados utilizando *embeddings* pré-treinados (Ali et al., 2025; Miric et al., 2023). A partir dessas representações, tornou-se possível aplicar métricas de comparação vetorial, como a similaridade do cosseno, a fim de medir a proximidade semântica entre termos. Essa etapa é fundamental em tarefas como expansão de vocabulário, recuperação de informação e análise de proximidade conceitual em corpora especializados, como textos de patentes (Teng et al., 2024). O presente experimento justifica-se, portanto, pela necessidade de superar as limitações de abordagens estritamente lexicais, viabilizando a identificação de sinônimos, significados

próximos e relações conceituais entre termos das patentes e a ontologia. A utilização da similaridade do cosseno permite mapear patentes a elementos ontológicos não apenas por coincidência literal, mas também por proximidade semântica, contornando questões relacionadas à desatualização dos termos da TRIZ (Berdyugina & Cavallucci, 2022a; Liang et al., 2008; Trapp & Warschat, 2025).

A **terceira alternativa experimental** envolve o emprego de LLMs para a geração de novos termos e padrões linguísticos, explorando relações previamente identificadas no espaço vetorial (Althammer et al., 2021; Trapp & Warschat, 2025). Essa integração entre pré-processamento, representação semântica e modelos generativos amplia significativamente as possibilidades de enriquecimento lexical, mostrando-se particularmente relevante em aplicações que demandam a expansão de ontologias e o aprimoramento de sistemas de mineração de texto (Choi, Park, et al., 2012; Russo, 2011; Souili, Cavallucci, & Rousselot, 2015b; Spreafico & Spreafico, 2021; Trapp & Warschat, 2025).

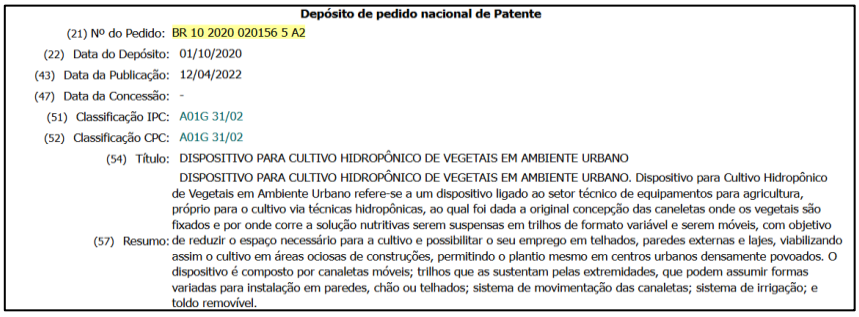
6.3.2 Obtenção dos Documentos de Patente

Os documentos de patente que compõem o *corpus* de análise foram obtidos no site do Instituto Nacional da Propriedade Industrial (INPI) do Brasil, por meio de um processo de *crawling*. O critério de extração foi aleatório, priorizando documentos publicados entre 2020 e 2024, em ordem decrescente temporal, sem restrições quanto ao campo tecnológico. Foram extraídos os campos textuais “[54] Título” e “[57] Resumo”, além dos campos estruturados “[21] Número do Pedido”, “[22] Data do Pedido” e “[51] Classificação IPC”. Estes últimos foram utilizados para caracterizar o perfil do *corpus* analisado.

Cabe destacar que os dados textuais disponibilizados pelo INPI se restringem ao título e ao resumo, uma vez que a descrição técnica e as reivindicações estão acessíveis apenas em documentos digitalizados no formato PDF, não sendo passíveis de extração automática. A Figura 21 apresenta um exemplo dos campos textuais disponíveis no site do INPI.

Figura 21

Exemplo dos campos textuais disponíveis para extração automática no site do Instituto Nacional de Propriedade Industrial do Brasil



Depósito de pedido nacional de Patente

(21) Nº do Pedido: BR 10 2020 020156 5 A2
(22) Data do Depósito: 01/10/2020
(43) Data da Publicação: 12/04/2022
(47) Data da Concessão: -
(51) Classificação IPC: A01G 31/02
(52) Classificação CPC: A01G 31/02
(54) Título: DISPOSITIVO PARA CULTIVO HIDROPÔNICO DE VEGETAIS EM AMBIENTE URBANO

(57) Resumo: DISPOSITIVO PARA CULTIVO HIDROPÔNICO DE VEGETAIS EM AMBIENTE URBANO. Dispositivo para Cultivo Hidropônico de Vegetais em Ambiente Urbano refere-se a um dispositivo ligado ao setor técnico de equipamentos para agricultura, próprio para o cultivo via técnicas hidropônicas, ao qual foi dada a original concepção das caneletas onde os vegetais são fixados e por onde corre a solução nutritivas serem suspensas em trilhos de formato variável e serem móveis, com objetivo de reduzir o espaço necessário para a cultivo e possibilitar o seu emprego em telhados, paredes externas e lajes, viabilizando assim o cultivo em áreas ociosas de construções, permitindo o plantio mesmo em centros urbanos densamente povoados. O dispositivo é composto por canaletas móveis; trilhos que as sustentam pelas extremidades, que podem assumir formas variadas para instalação em paredes, chão ou telhados; sistema de movimentação das canaletas; sistema de irrigação; e toldo removível.

Nota. Obtido no INPI (2025).

Os dados extraídos foram organizados em uma planilha eletrônica para posterior processamento e análise.

6.3.3 Avaliação dos Resultados

Para a avaliação do desempenho do método de mineração textual apoiado por uma ontologia baseada nos efeitos físicos da TRIZ, são previstos dois momentos de especial importância: (i) o processo de construção do formulário de avaliação, que inclui o objetivo a ser alcançado, a definição da escala de avaliação, a elaboração das quesões e a verificação da consistência interna por meio da validação da confiabilidade estatística e (ii) o processo de seleção dos especialistas.

6.3.3.1 Construção do instrumento de avaliação

O objetivo do processo avaliativo conduzido por especialistas é analisar o desempenho do método de mineração textual quanto à consistência dos componentes semânticos extraídos (Tarefa (T), Objeto (O) e Efeito Físico (EF)) em relação ao conteúdo do título e do resumo da patente, bem como avaliar a coerência dos componentes semânticos gerados pelo método em comparação ao registro da ontologia ao qual estão vinculados.

As respostas dos especialistas ao formulário de avaliação permitem mensurar a taxa de acerto na identificação de relações semanticamente consistentes, bem como a taxa de erro decorrente da classificação incorreta de relacionamentos inconsistentes como consistentes. Essas métricas possibilitam avaliar a precisão e a confiabilidade do processo de extração dos componentes semânticos T, O e EF.

Para a construção do instrumento de avaliação, foram considerados critérios como o tempo de resposta do avaliador, a objetividade das questões, a eliminação de ambiguidades e a

pertinência das perguntas em relação ao objetivo da avaliação (Rivera-Garrido et al., 2022). Nesse contexto, elaborou-se um formulário eletrônico na plataforma *Google Forms*, escolhida por possibilitar uma apresentação clara e organizada dos campos textuais e das questões — aspecto particularmente relevante, dada a necessidade de consultas frequentes ao texto das patentes para confirmação das respostas.

Considerando o tempo estimado de preenchimento, de até cinco horas, a ferramenta foi também selecionada por permitir o salvamento automático das respostas, possibilitando que o avaliador retomasse o preenchimento a qualquer momento, sem risco de perda de dados.

Foram previstas quatro questões com resposta única, organizadas em escalas ordinalmente codificadas de forma consistente, de modo que valores mais baixos indiquem maior compatibilidade (1 = compatibilidade alta; 2 = compatibilidade inexistente).

As **Questões 1 e 2** utilizam uma escala de quatro pontos. Em cada questão são previstas quatro possibilidades de resposta única obrigatória, onde os itens 1 e 2 correspondem a uma escala dicotômica (sim/não). Já os itens 3 e 4 são opções complementares, assinaladas somente quando não é possível responder com 1 ou 2. O item 3 indica insuficiência de informações no campo textual analisado; o item 4 indica que os termos Tarefa (T) e Objeto (O) são genéricos e pouco representativos em relação ao conteúdo textual. As questões 1 e 2 são a seguir apresentadas:

Q1 - Compatibilidade com o título
As subclasses Tarefa (T) e Objeto (O) atribuídas à patente são compatíveis com o título?
(1) Sim — T e/ou O aparecem explicitamente ou de forma inferível no campo textual analisado.
(2) Não — T e/ou O não têm relação com o campo textual analisado e não são inferíveis.
(3) Indeterminado — informações do campo textual analisado são insuficientes.
(4) Indeterminado — o relacionamento T–O é demasiado genérico.

Q2 - Compatibilidade com o resumo
As subclasses Tarefa (T) e Objeto (O) atribuídas à patente são compatíveis com o resumo?
(1) Sim — T e/ou O aparecem explicitamente ou de forma inferível no campo textual analisado.
(2) Não — T e/ou O não têm relação com o campo textual analisado e não são inferíveis.
(3) Indeterminado — informações do campo textual analisado são insuficientes.
(4) Indeterminado — o relacionamento T–O é demasiado genérico.

As Questões 3 e 4 apresentam uma escala binária (sim/não), alinhada à direção conceitual das questões anteriores (1 = compatibilidade alta; 2 = compatibilidade inexistente). A questão 3 é a seguir apresentada.

Q3 - Relacionamento ternário
O relacionamento ternário atribuído à patente fornece uma sugestão de solução técnica

que é efetivamente abordada ou resolvida pela patente?
(1) Sim
(2) Não

Q4 - Existe consistência entre os termos derivados (T, O, EF) e aqueles dos quais se originam?
(1) Sim
(2) Não

As Questões 1 e 2 têm como objetivo mensurar a precisão, por meio de métricas quantitativas, e o poder representacional dos termos atribuídos à patente. Neste contexto, entende-se por poder representacional a proximidade semântica entre os termos T e O, atribuídos pelo método de mineração textual, e o conteúdo informacional do título e do resumo.

As Questões 3 e 4, por sua vez, buscam analisar se os termos T, O e Efeito Físico (EF) atribuídos fornecem uma lógica contextual coerente com o conteúdo textual do título e do resumo da patente analisada e relacionados com os termos originais da ontologia, no caso de derivação.

A escolha de uma escala binária se deveu ao fato de não se pretender uma variedade de valores obtida por meio de escalas múltiplas, como a de Likert. O objetivo central foi identificar se o método de mineração textual atribui termos da ontologia (ou termos derivados dela, por meio de enriquecimento) que de fato representam o conteúdo informacional do campo da patente, bem como verificar se o relacionamento entre T, O e EF reflete uma potencial solução genérica (solução conceitual), alinhada à abordagem sistemática de resolução de problemas utilizada na TRIZ (Gadd, 2011; Savransky, 2000).

Dessa forma, a escala utilizada nas quatro questões do formulário de avaliação permite identificar um VP, onde o relacionamento semântico atribuído pelo modelo é confirmado ou considerado consistente pelo especialista (portanto, é atribuído "(1) Sim"), ou FP, onde o relacionamento semântico atribuído pelo modelo é considerado inconsistente (atribuído “(2) Não”) ou inválido (atribuído “(3)” ou “(4)”).

O questionário de respostas foi submetido a um teste piloto com três especialistas, sendo dois bibliotecários com especialidade em tecnologia da informação e um engenheiro, todos familizarizados com patentes de invenção, a fim de revisar possíveis ambiguidades. Nessa etapa, os especialistas sugeriram incluir o conceito de “compatibilidade”, empregado nas questões Q1 e Q2, nas instruções de preenchimento, além de revisar a redação dos itens 1, 2 e 3 dessas questões. Foi ainda acrescentada, embora de forma redundante, a expressão “campo textual analisado” para maior clareza interpretativa.

6.3.3.2 Processo de seleção dos especialistas

Para validar os resultados obtidos pelo método de mineração, os termos extraídos dos campos textuais de título e resumo de documentos de patente extraídos do site do INPI foram avaliados por especialistas de diferentes áreas do conhecimento, com formação interdisciplinar que contempla tanto domínio técnico específico quanto familiaridade com documentos de patente.

A seleção ocorreu por convite direto e o processo avaliativo foi detalhadamente formalizado (Apêndice E), incluindo uma reunião para apresentação da ontologia e dos conceitos da TRIZ. Para preservar a imparcialidade e mitigar vieses associados a fatores pessoais ou profissionais, o anonimato dos participantes foi estritamente mantido através de codificação.

O grupo de avaliadores foi composto por especialistas de diferentes áreas do conhecimento: três em Farmácia, dois em Engenharia Mecânica, um em Engenharia Elétrica e um em Química. A seguir, a Tabela 17 apresenta o perfil dos participantes, conforme a formação e o respectivo código de identificação.

Tabela 17

Formação dos especialistas

Código do Especialista	Formação
E1	Engenharia mecânica
E2	Engenharia mecânica
E3	Engenharia elétrica
E4	Farmácia
E5	Farmácia
E6	Farmácia
E7	Química

Nota. Dados da pesquisa (2025).

Para a avaliação da extração de termos executada pelo método, cada especialista recebeu um formulário *online* (Apêndice F) contendo as instruções de preenchimento e uma relação de documentos de patente, com número do depósito, título, resumo e os termos da ontologia extraídos pelo método ou atribuídos e vinculados a um relacionamento T-O-EF na ontologia. O objetivo do processo avaliativo realizado pelos especialistas é validar os resultados da extração executada pelo método de mineração de inteligência técnica, concordando ou discordando da qualidade do processo de extração de termos.

6.4 Resultados

Esta seção apresenta, de forma detalhada, as etapas do processo de mineração textual conduzido no estudo. São descritos o procedimento de obtenção dos documentos de patente e as alternativas experimentais inicialmente definidas, juntamente com os resultados obtidos, culminando na avaliação do desempenho do método desenvolvido, realizada com o apoio de especialistas.

6.4.1 Obtenção e Estruturação dos Documentos de Patente

Foram obtidos 3.299 documentos de patente na base do INPI, considerando depósitos realizados a partir do ano de 2020. A amostragem foi conduzida de forma aleatória, sem delimitação quanto ao campo tecnológico, uma vez que o objetivo consiste em avaliar a performance do método em domínios distintos do conhecimento.

Os documentos selecionados foram organizados em uma planilha eletrônica (*Microsoft Excel*), com campos predefinidos para as seguintes informações: número da patente, data de depósito, Classificação IPC, título, resumo e localizador de recursos (URL). A Figura 22 ilustra a estrutura da planilha eletrônica e os respectivos campos.

Figura 22

Estrutura dos dados bibliográficos e textuais dos documentos de patente em planilha eletrônica

pedido	data_deposito	titulo	ipc	url	resumo	classifica_ipc
BR 11 2021 18393 0	02/03/2020	TRATAMENTO DE COLISÕES EM UPLINK	H04L 1/18	https://busca.inpi.gov.br/pePI/servlet/Patente...	A presente invenção se refere a métodos, sis...	H04L 1/18
BR 11 2021 18071 0	02/03/2020	ALOJAMENTO DE VELA DE IGNIÇÃO COM PROTEÇÃO ANT...	H01T 13/14	https://busca.inpi.gov.br/pePI/servlet/Patente...	ALOJAMENTO DE VELA DE IGNIÇÃO COM PROTEÇÃO A...	H01T 13/14 ; H01T 13/20 ; H01T 13/32 ; H0...
BR 11 2021 16947 4	02/03/2020	ANTICORPOS QUE RECONHECEM TAU	C07K 16/18	https://busca.inpi.gov.br/pePI/servlet/Patente...	ANTICORPOS QUE RECONHECEM TAU. A invenção fo...	C07K 16/18 ; G01N 33/68

Nota. Dados da pesquisa (2025).

Os campos bibliográficos e textuais foram mantidos integralmente, conforme extraídos da base do INPI, para subsidiar as etapas subsequentes do processo de pré-processamento.

6.4.2 Primeira Alternativa Experimental: Marcação Morfossintática (PoS Tagging)

Na etapa de pré-processamento, os textos correspondentes ao título e ao resumo (texto de entrada) foram submetidos a tratamento utilizando a biblioteca *SpaCy*. Esse procedimento envolveu: (i) conversão integral para caracteres minúsculos; (ii) lematização das palavras, reduzindo-as à sua forma base; (iii) exclusão das stopwords; e (iv) remoção de sinais de pontuação e de acentuação. A Figura 23 apresenta um exemplo desse processo, comparando a versão original dos textos com a versão resultante após a lematização.

Figura 23

Exemplo da transformação de título e resumo de documentos de patente: versão original *versus* versão lematizada

	titulo	titulo_lemmatized
0	TRATAMENTO DE COLISÕES EM UPLINK	tratamento colisao uplink
1	ALOJAMENTO DE VELA DE IGNIÇÃO COM PROTEÇÃO ANT...	alojamento vela ignicao protecao anticorrosivo...
2	ANTICORPOS QUE RECONHECEM TAU	anticorpo reconhecer tau
3	AQUECEDOR DE AR A LENHA COM DUPLA EXAUSTÃO PAR...	aquecedor ar lenha dupla exaustao utilizar amb...
4	BIBLIOTECAS DE CÉLULAS ÚNICAS E NÚCLEOS ÚNICOS...	biblioteca celula unico nucleo unico alto rend...
5	CAPACIDADE DE MONITORAMENTO DE CANAL DE CONTRO...	capacidade monitoramento canal controle comuni...
6	CÉLULA, CONSTRUÇÃO DE EXPRESSÃO, PLANTA OU PAR...	celula construcao expressao planta planta conj...

	resumo	resumo_lemmatized
0	A presente invenção se refere a métodos, sis...	presente invencao referir metodo sistema di...
1	ALOJAMENTO DE VELA DE IGNIÇÃO COM PROTEÇÃO A...	alojamento vela ignicao protecao anticorros...
2	ANTICORPOS QUE RECONHECEM TAU. A invenção fo...	anticorpo reconhecer tau invencao fornecer ...
3	AQUECEDOR DE AR A LENHA COM DUPLA EXAUSTAO P...	aquecedor ar lenha dupla exaustao utilizar ...
4	BIBLIOTECAS DE CÉLULAS ÚNICAS E NÚCLEOS ÚNIC...	biblioteca celula unico nucleo unico alto r...
5	CAPACIDADE DE MONITORAMENTO DE CANAL DE CONT...	capacidade monitoramento canal controle com...
6	CÉLULA, CONSTRUÇÃO DE EXPRESSÃO, PLANTA OU P...	celula construcao expressao planta planta c...

Nota. Dados da pesquisa (2025).

Os campos textuais de título e de resumo foram reduzidos para seguirem à etapa posterior de transformação em informações estruturadas, facilitando o processo de mineração.

Os textos previamente normalizados dos títulos e resumos foram submetidos à marcação morfossintática (PoS *tagging*) por meio da biblioteca *SpaCy*. Nesse processo, cada termo do texto foi identificado e classificado de acordo com sua respectiva categoria morfossintática, como substantivo, verbo, adjetivo, entre outras, o que possibilita análises linguísticas mais precisas e a extração de padrões relevantes para etapas posteriores do estudo. Na Figura 24 é

exemplificada a marcação morfossintática do título e do resumo de uma das patentes da amostra.

Figura 24

Exemplo da marcação morfossintática de título e resumo de documento de patente

```
Original text POS
metodo -> NOUN (NOUN)
codificacao -> ADJ (ADJ)
video -> PROPN (PROPN)
codificador -> ADJ (ADJ)
decodificador -> ADJ (ADJ)
produto -> NOUN (NOUN)
programa -> ADJ (ADJ)
computador -> ADJ (ADJ)
```

```
Processed text POS
MÉTODO -> PROPN (PROPN)
DE -> ADP (ADP)
CODIFICAÇÃO -> PROPN (PROPN)
DE -> PROPN (PROPN)
VÍDEO -> PROPN (PROPN)
, -> PUNCT (PUNCT)
CODIFICADOR -> PROPN (PROPN)
, -> PUNCT (PUNCT)
DECODIFICADOR -> PROPN (PROPN)
E -> CCONJ (CCONJ)
PRODUTO -> PROPN (PROPN)
DE -> ADP (ADP)
PROGRAMA -> PROPN (PROPN)
DE -> ADP (ADP)
COMPUTADOR -> NOUN (NOUN)
```

Nota. Dados da pesquisa (2025).

Com base nos resultados obtidos tanto no trecho de código quanto na saída do *SpaCy*, conclui-se que a marcação *PoS* não foi eficaz para identificar corretamente os papéis gramaticais das palavras no texto.

Durante o processo de lematização, observou-se que várias palavras foram rotuladas de forma inadequada, comprometendo a interpretação linguística esperada. Exemplos incluem:

- “codificacao” → ADJ (classificado como adjetivo, deveria ser um substantivo);
- “programa” → ADJ (classificado como adjetivo, deveria ser um substantivo);
- “metodo” → NOUN (classificado corretamente como substantivo, porém perdeu a relação com o verbo associado).

No texto processado (não lematizado), verificou-se que a maioria dos termos relevantes foi classificada como PROPN (nome próprio), o que distorce significativamente o objetivo da análise. Embora a biblioteca *spaCy* disponibilize uma marcação *PoS* própria para o português, por se tratar de uma língua com poucos recursos, o modelo mostra-se menos competente na marcação morfossintática. Além disso, os textos de patentes apresentam irregularidades na formatação, com o uso alternado de letras maiúsculas e minúsculas, o que reduziu a eficiência do modelo na identificação de nomes próprios e outras classes sintáticas. Essa distorção inviabiliza a identificação adequada dos elementos principais pretendidos:

- Tarefa: representada por verbos como “codificar”, “decodificar”;

- Objetos: representados por substantivos como “vídeo”, “produto”, “computador”.

Em síntese, a marcação automática de *PoS* pelo modelo do *SpaCy* apresentou limitações expressivas na correta distinção de classes gramaticais, comprometendo a extração semântica necessária à identificação de tarefas e objetos nas patentes analisadas.

6.4.3 Segunda Alternativa Experimental: Representação Vetorial

Para superar as limitações observadas na marcação morfossintática (*PoS tagging*) do *SpaCy* em textos técnicos em português, foi implementada uma abordagem alternativa baseada na distância do cosseno entre vetores semânticos, com o objetivo de identificar Tarefas e Objetos por meio da comparação vetorial que quantifica relações semânticas entre termos.

Assim como na primeira alternativa experimental, os textos referentes ao título e ao resumo (texto de entrada) foram pré-processados utilizando a biblioteca *SpaCy*. O procedimento contemplou: (i) conversão integral para caracteres minúsculos; (ii) lematização das palavras, reduzindo as palavras a sua forma base; (iii) exclusão das stopwords; e (iv) remoção de sinais de pontuação e de acentuação.

Em seguida, os termos da ontologia, bem como o título e o resumo das patentes, foram vetorizados, ou seja, transformados em representações numéricas — chamadas vetores — que podem ser manipuladas matematicamente por algoritmos de PLN. Para essa vetorização, foi utilizado o modelo treinado pela biblioteca *SpaCy* para a língua portuguesa, disponível em [https://SpaCy.io/models/pt#pt_core_news_lg] e [https://github.com/explosion/SpaCy-models/releases/tag/pt_core_news_lg-3.8.0]. Cada termo foi convertido em um vetor de 300 dimensões, ou seja, uma sequência de 300 números reais que mantêm relações sintáticas entre as diferentes palavras do léxico da palavra no espaço vetorial.

Esse modelo é classificado como “large” (lg), o que significa que possui um vocabulário mais extenso e vetores de representação de maior dimensão quando comparado aos modelos “sm” (*small*) e “md” (*medium*). É adequado para tarefas de PLN em português, oferecendo desempenho mais robusto graças ao suporte de vetores semânticos de alta qualidade. Além disso, inclui componentes como vocabulário, análise sintática (dependência e árvore sintática), reconhecimento de entidades nomeadas (NER) e vetores de palavras, que permitem o cálculo de similaridades semânticas entre termos.

A função *find_matches* foi utilizada para decompor a *string* de entrada em unidades lexicais e compará-las com vetores pré-calculados (no caso, representações das tarefas e dos objetos). A identificação pôde ser conduzida a partir de dois critérios: (i) aplicação de um limiar

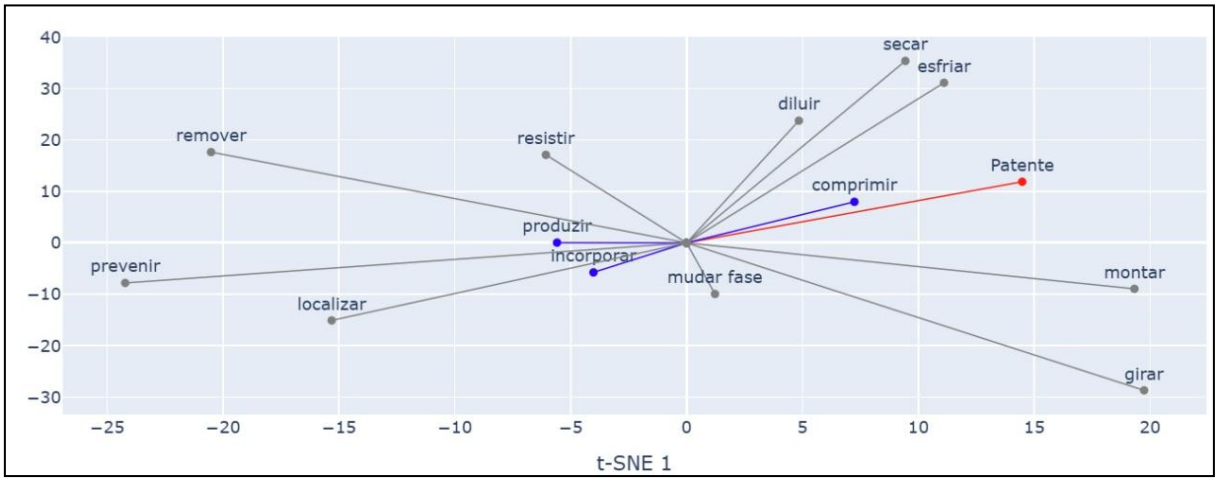
de similaridade (*threshold*), ou (ii) seleção dos top_n resultados mais próximos.

Para ilustrar como esses vetores preservam suas características léxicas, a patente intitulada “Método de codificação de vídeo, codificador, decodificador e produto de programa de computador” foi pré-processada, resultando em “metodo codificacao video codificador decodificador produto programa computador” após a lematização e a remoção de *stopwords*. Em seguida, o texto foi vetorizado e comparado à lista de tarefas da ontologia TRIZ, destacando-se as três tarefas mais próximas, utilizando a similaridade de cosseno, e exibindo também outras tarefas que não figuraram entre as mais próximas.

Para que os vetores pudessem ser representados em duas dimensões (uma vez que originalmente possuem 300 dimensões), empregou-se a técnica de redução de dimensionalidade t-SNE (*t-distributed Stochastic Neighbor Embedding*) (Maatens & Hinton, 2008). Essa técnica permite visualizar dados de alta dimensionalidade, atribuindo a cada ponto de dado uma localização em um mapa bidimensional ou tridimensional. A Figura 25 apresenta o mapa bidimensional dos vetores de tarefas da ontologia TRIZ, bem como os vetores mais próximos aos vetores correspondentes ao título da patente.

Figura 25

Exemplo da vetorização de título de documentos de patente



Nota. Dados da pesquisa (2025).

No entanto, em determinados casos, observou-se a ocorrência do aviso *RuntimeWarning* (divisão por zero). Esse comportamento resulta da presença de vetores com norma nula (*np.linalg.norm(vector) == 0*), o que inviabiliza o cálculo da similaridade devido à divisão por zero. Tal situação decorre do fato de o *SpaCy* não gerar representações vetoriais para certos termos, em função das limitações do vocabulário contemplado no modelo pré-treinado.

O modelo treinado para língua portuguesa e utilizado nesse experimento, mesmo sendo um modelo “grande”, não garantiu precisão para textos de patentes, especialmente em termos

especializados que não estavam presentes no conjunto de treinamento. De toda forma, ao comparar *embeddings* de palavras com vetores pré-calculados, foi possível identificar candidatos a Tarefas e Objetos com base na proximidade semântica, mas não resolveu completamente os desafios de classificação correta dos papéis gramaticais.

6.4.4 Terceira Alternativa Experimental: Modelos de Linguagem de Grande Escala

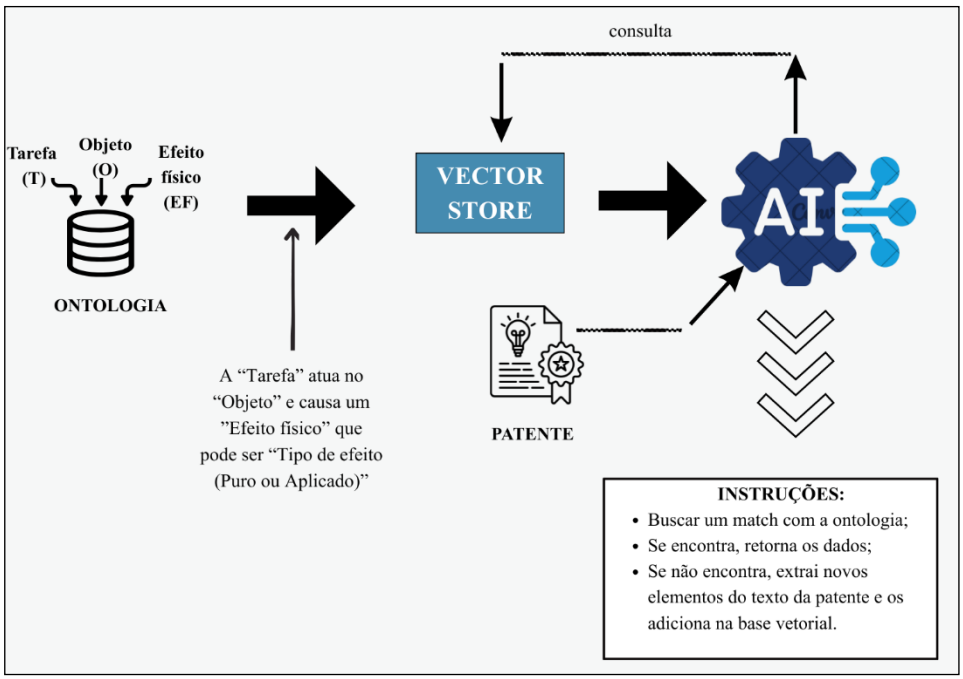
Os LLMs, como o GPT-4, representam um avanço significativo nas tarefas de interpretação textual, devido à sua capacidade de lidar com contextos amplos e estabelecer conexões semânticas complexas (OpenAI, 2023). Segundo Trapp e Warschat (2025), a integração de LLMs aos processos de extração de conceitos técnicos aumenta significativamente a eficiência da análise de como patentes quando comparada aos métodos tradicionais, como a PNL. Os LLMs reduzem a necessidade de ajustes finos extensivos, lidam com grandes volumes de dados de forma mais econômica e melhoram a qualidade do pré-processamento, tornando-os uma alternativa poderosa às metodologias tradicionais, muitas vezes complicadas, de vários estágios (Jiang et al., 2025; Trapp & Warschat, 2025).

6.4.4.1 Arquitetura exclusiva

Inicialmente foi utilizada uma arquitetura exclusiva, utilizando-se um LLM puro, sem integração com componentes externos. O modelo opera apenas com seu conhecimento pré-treinado, sem recorrer a ferramentas auxiliares e nem bases de dados externas, ou especializações modulares. O uso dessa arquitetura foi justificado para avaliar até que ponto o modelo poderia, de forma autônoma, identificar e correlacionar elementos das patentes com os registros da ontologia TRIZ, a partir de um repositório vetorial. Esse teste buscou verificar a capacidade da IA de atuar como um sistema de mapeamento direto, sem a necessidade de etapas adicionais de decisão. Na Figura 26 é apresentada a representação esquemática da arquitetura exclusiva.

Figura 26

Representação esquemática da arquitetura exclusiva



Nota. Elaborado pela Autora (2025).

Em um primeiro momento, cada registro da ontologia é convertido em uma frase seguindo o modelo: "A {Tarefa} atua no {Objeto} e causa um {Efeito físico}, que pode ser um {Tipo de Efeito}." Essa transformação permite que cada registro seja armazenado em um formato textual interpretável pela IA.

A partir dessa conversão, cada entrada da ontologia é transformada em uma frase e representada como um vetor em uma base de dados vetorial — neste caso, utilizando o ChromaDB como *Vector store*. Essa estrutura vetorial possibilita que as frases sejam recuperadas e utilizadas na extração dos elementos técnicos presentes nas patentes.

A base vetorial torna-se acessível à IA generativa por meio de um processo de recuperação de informação (*retrieval*). Antes de gerar uma resposta, a IA consulta essa base de conhecimento vetorial e compõe a resposta utilizando os registros mais relevantes, conforme uma lógica de proximidade vetorial entre a pergunta e os vetores armazenados. O processo pode ser descrito da seguinte forma:

(1) O sistema recebe o título/resumo de uma patente, que terá seus metadados minerados;

(2) O texto recebido é vetorizado;

(3) O vetor correspondente é comparado com os vetores da base, e os três mais similares são recuperados;

(4) Com base nesses elementos similares e nas instruções embutidas no *prompt*, a IA avalia se há algum registro que descreva adequadamente a patente recebida ou se é necessário extrair novos elementos do texto da patente;

(5) Os elementos identificados são retornados como resposta;

(6) Caso novos elementos tenham sido gerados, eles são adicionados à base vetorial, expandindo a Ontologia existente e aprimorando continuamente o modelo.

Dois hiperparâmetros foram ajustados para identificar o melhor arranjo do sistema: o número de elementos retornados pela base vetorial e a temperatura das gerações.

- Número de elementos retornados: 3 (refere-se à quantidade de vetores mais próximos retornados pela base).
- Temperatura das gerações: zero (parâmetro que controla o grau de aleatoriedade na geração das respostas. Valores mais baixos resultam em respostas mais determinísticas e consistentes, enquanto valores mais altos aumentam a criatividade, mas também o risco de incoerências).

O *prompt* a seguir foi desenvolvido com o objetivo de orientar o modelo de inteligência artificial na execução da mineração semântica de patentes, conforme os princípios estabelecidos pela ontologia da TRIZ. O documento define de maneira estruturada os parâmetros de extração e o formato esperado dos resultados, assegurando rigor metodológico e coerência na análise semântica realizada pelo sistema.

Você receberá o título e o resumo de uma patente em português.
*A partir destes dados, vamos buscar minerar dados desta patente seguindo a **Ontologia da TRIZ**.*
Esta metodologia busca descrever uma patente a partir de uma série de elementos. Sua tarefa é identificar e retornar os seguintes elementos:
- `task`: ação ou tarefa (ex: "Aquecer");
- `object`: objeto-alvo da ação (ex: "Água");
- `kind_effect`: tipo de efeito causado no objeto a partir da tarefa (ex: "Aumento da temperatura");
-`physical_effect`: o efeito físico utilizado ou observado (ex: "Efeito Joule").
É importante notar que a mineração dos dados não busca extrair exatamente os componentes da patente, mas sim um conjunto de termos e características suficientemente genéricos que consigam descrever a patente.
Além da patente e os metadados extraídos, você receberá um conjunto de dados já existentes na TRIZ que podem ser usados para descrever a patente. Sua tarefa será decidir se há algum conjunto de dados já existentes da TRIZ que podem ser usados para descrever a patente ou se os dados minerados serão utilizados. É importante tentar priorizar um elemento que já exista na TRIZ. Só traga um novo elemento se realmente nenhum elemento se aproximar do contexto.
Em resumo, a lógica adotada será:
1. Sabendo que devo priorizar elementos já existentes na TRIZ, devo descrever esta patente com os dados extraídos ou usar algum conjunto de dados da TRIZ que recebi;

2. Caso eu utilize algum elemento já existente, devo retornar o mesmo com o `derived_from` nulo;
3. Caso eu utilize o elemento que foi fornecido, devo retornar o mesmo e apontar o elemento já existente na TRIZ que mais se aproxime a partir do `derived_from`.
4. Neste caso, `derived_from` recebe o `id` dos metadados do elemento mais próximo
Regras:
- Todos os campos de texto são obrigatórios
- O campo `derived_from` deve ser:
- `null` se o elemento já existe na base TRIZ
- Uma string com o `id` do elemento mais próximo existente se for derivado.
- NÃO INVENTE o id. Ele deve ser algum dos ids dos metadados listados em seu input
- Mantenha a terminologia consistente com os elementos TRIZ existentes
- Use descrições claras e específicas para cada campo
 # Formato de saída
O formato de saída deve ser um JSON com os seguintes campos:
- `derived_from`: `id` do elemento mais próximo já existente na TRIZ.
- `effect`: tipo de efeito (ex: "Aumento da temperatura");
- `task`: `task` Tarefa (ex: "Aumentar a temperatura");
- `object`: `object` Objeto (ex: "Água");
- `physical_effect`: `physical_effect` Efeito físico (ex: "Aumento da temperatura").
Exemplo:
```json
{{“derived_from": "81071a34-d099-4dca-b07e-aaf120297f61", // Opcional: id da origem do registro
"effect": "Efeito", // Tipo do efeito (ex: "Mecânico", "Térmico")
"task": "Identificação" // Ação/tarefa (ex: "Comprimir", "Aquecer")
"object": "Gás", // Objeto alvo (ex: "Água", "Metal")
"physical_effect": "Adição de massa" // Efeito físico (ex: "Aumento de temperatura")
}}

Nesta experimentação, o principal problema identificado foi a dificuldade em alcançar um equilíbrio na etapa generativa. O modelo foi sobrecarregado com uma tarefa complexa, que exigia, em uma única operação, o desempenho de duas competências distintas:

(a) Avaliação Crítica e Comparação: Analisar o termo extraído e compará-lo com os elementos já existentes na Ontologia (função de recuperação e *matching*).

(b) Decisão e Geração: Decidir se deveria usar um elemento existente ou gerar um novo termo semântico ou relacionamento ontológico (função generativa e de derivação).

O acúmulo dessas responsabilidades aparentemente comprometeu o desempenho do modelo, o que está em conformidade com as boas práticas de engenharia de software e inteligência artificial. Ao se exigir que o modelo realizasse simultaneamente as tarefas de avaliação e geração, houve perda de desempenho em ambas as etapas, resultando em um equilíbrio insatisfatório e baixa precisão na avaliação dos elementos.

Na Tabela 18 são apresentados casos ilustrativos de extrações de subclasses da ontologia, nos quais se observa falta de consistência entre os termos sugeridos e o conteúdo

textual do título e do resumo analisado. Em vários casos, o modelo recuperou termos com base apenas em similaridade léxica, sem considerar adequadamente o contexto semântico.

Tabela 18

Casos ilustrativos de extrações de termos do conteúdo textual da patente com atribuição de termos da ontologia utilizando Modelos de Linguagem de Grande Escala

Título	Transcrição do resumo	Subclasses extraídas da ontologia
Equipamento para preparação customizada e venda de produtos	Equipamento (100) para preparação customizada e venda de produtos que objetiva possibilitar ao usuário (consumidor) selecionar, dosar e misturar diversos tipos de sabores, essências e (ou) cores para assim definir um produto (P) em especial. O equipamento (100) aqui tratado inclui: a) um sistema computacional; b) um sistema de armazenamento das substâncias (fluidos) com as quais os produtos serão preparados; c) um sistema de pressurização das substâncias com as quais os produtos serão preparados; d) um sistema de dosagem seletiva das substâncias com as quais os produtos serão preparados; e) um sistema de armazenamento e movimentação de funis utilizados no processamento dos produtos; f) um sistema de armazenamento e movimentação dos recipientes nos quais os produtos serão envasados; g) um sistema de pesagem; h) um sistema de alimentação de tampas para o fechamento dos recipientes; i) um sistema de aplicação das tampas aos recipientes; j) um sistema de movimentação e apresentação ao consumidor dos recipientes já envasados; k) um sistema de refrigeração; l) uma unidade de impressão de rótulos; e m) uma tela LCD sensível ao toque como interface gráfica para operação do usuário.	Tipo de Efeito - Aplicação Tarefa - Aperta Objeto - Líquido Efeito físico - Materiais refratários
Processo de fabricação de base para bancada em poliestireno expansível e base para bancada em poliestireno expansível.	Trata de um processo que utiliza pequenos blocos (2) de EPS expandido colados com adesivo de EPS e flocos de EPS na proporção de 2x1 entre a face superior e inferior de uma tela metálica (4), responsável por reforçar a região de instalação da cuba (5) e da torneira (6), de forma que a base (1) para bancada de EPS obtida é leve, robusta e de fácil instalação	Tipo de efeito - Aplicação Tarefa - Expandir Objeto - Sólido Efeito físico - Espuma metálica
Sistema de preparação de bebidas aromáticas, como	A presente invenção revela um sistema de preparação de bebidas aromáticas, como por exemplo chá, compreendendo um aparelho (1) apresentando uma superfície de topo que	Tipo de tarefa - Aplicação Tarefa - Separar Objeto - Líquido

Título	Transcrição do resumo	Subclasses extraídas da ontologia
por exemplo chá	compreende uma disposição de colocação de recipiente (9) adaptada para colocação de um recipiente de fluido de preparação (11, 11?), e uma região de serviço adaptada de modo que proporciona espaço para colocação de uma pluralidade de recipientes de bebida (10, 10?), sendo que o referido aparelho (1) apresenta uma altura (H) que é pelo menos quatro vezes menor que qualquer das outras duas dimensões características (D2, D3) e a superfície de topo se desenvolve desprovida de qualquer projeção ou reentrância numa área pelo menos aproximadamente similar à área da região de base e maior que a área das superfícies frontal, posterior e laterais.	Efeito físico - Destilação a vácuo

Nota. Dados da pesquisa (2025).

O resultado insatisfatório sugere que a arquitetura do método precisa ser refinada, seguindo o princípio da ou modularidade na IA. A solução implícita para esse tipo de problema é desmembrar o processo em etapas distintas:

(1) Recuperação e Classificação: uma parte do modelo deve ser responsável por avaliar se o termo extraído tem um correspondente satisfatório na ontologia existente.

(2) Derivação e Validação: somente se a etapa 1 não fornecer um *match* adequado, o sistema deve acionar o componente generativo para criar novos elementos.

Essa modularização permite que cada componente do método híbrido exerça sua competência de forma adequada (interpretação semântica profunda, validação lógica estruturada e adaptação ontológica dinâmica), potencializando a precisão do método final.

6.4.4.2 Arquitetura híbrida

Com base nos resultados obtidos, foi testada uma arquitetura híbrida que combina o uso de LLMs com módulos analíticos auxiliares, neste caso, um componente externo denominado Lógica de Decisão Derivativa, o qual opera segundo regras lógicas explícitas e ajustáveis. Nesse fluxo, o LLM realiza a extração inicial dos componentes semânticos Tarefa, Objeto e Efeito Físico, enquanto a decisão final sobre sua adoção ou associação a registros da TRIZ é delegada ao componente analítico. Conforme defendem Bender et al. (2021), a integração entre LLMs e bases de conhecimento estruturadas aumenta a confiabilidade e a consistência dos resultados, além de permitir a expansão controlada da ontologia.

Na arquitetura híbrida, utilizando o modelo GPT-4.1, o título e o resumo de cada patente foram processados em conjunto com os dados armazenados no *vector store*. A partir dessa

combinação, a IA deveria identificar os elementos ontológicos presentes na patente e decidir se deveria adotar o novo elemento extraído ou selecionar, por similaridade semântica, um dos candidatos disponíveis no *vector store*.

Especificamente, o LLM é responsável por extrair os componentes semânticos (Tarefa, Objeto e Efeito Físico) a partir do conteúdo textual da patente. Em seguida, o componente analítico de decisão avalia se os elementos extraídos devem ser mantidos como novos registros ou se já existe, na ontologia, algum elemento capaz de representar adequadamente a mesma relação.

Quando as extrações realizadas pela IA Generativa são validadas, elas são incorporadas à ontologia como novos registros, vinculados ao registro pré-existente mais semelhante, garantindo a consistência semântica e a rastreabilidade das novas inclusões.

Nesta etapa, os vetores gerados na seção anterior (arquitetura exclusiva) foram integrados a uma configuração desenvolvida com o *LangChain* — um *framework* de código aberto, disponível em *Python* e *JavaScript*, voltado ao desenvolvimento de aplicações baseadas em LLMs (como o GPT). O fluxo de processamento compreendeu as seguintes subetapas:

(1) Formatação da Patente

Cada documento de patente foi estruturado como uma string no formato:

Título: {{title}}

Resumo: {{abstract}}

(2) Extração TRIZ via LLM

O *TRIZExtractor* empregou um modelo de linguagem (por exemplo, o ChatGPT 4.1) para identificar e extrair elementos da ontologia a partir do texto da patente, utilizando temperatura 0.5 (zero ponto cinco). Dessa forma, o modelo indica um meio-termo entre o modo puramente determinístico e o modo altamente criativo, gerando uma saída com aleatoriedade moderada.

Os componentes extraídos incluíram:

- *task* (str): ação realizada (e.g., *"*Comprimir", "Aquecer");
- *object* (str): objeto alvo (e.g., *"*Água", "Metal");
- *physical_effect* (str): efeito físico resultante (e.g., "Aumento de temperatura").

(3) Busca por Similaridade Semântica

Os elementos TRIZ extraídos foram comparados às entradas existentes no repositório vetorial por meio de busca semântica, a fim de identificar os registros mais próximos já presentes na base de conhecimento.

(4) Lógica de Decisão Derivativa

O *TRIZExtractor* aplicou uma regra de decisão para evitar redundâncias nos registros:

- Caso o elemento extraído apresentasse diferença significativa em relação ao item mais próximo (com base em um limiar de similaridade definido como *threshold* 0.3), ele era registrado como um novo relacionamento, sendo adicionado ao repositório e vinculado ao item mais semelhante via campo *derived_from*;
- Caso não houvesse diferença relevante, o sistema retornava o elemento TRIZ existente, evitando duplicações.

O *threshold* é a régua de corte usada para determinar se a saída do modelo é aceitável/ consistente, selecionando apenas registros com pontuações abaixo de 0.3.

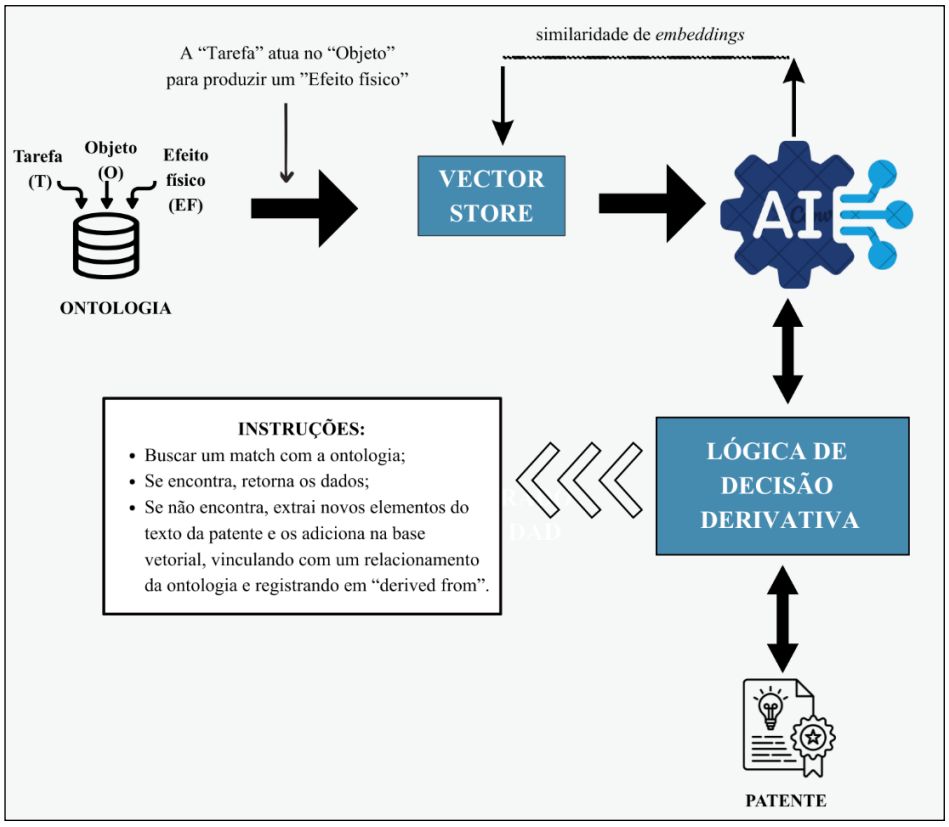
(5) Expansão da Base de Conhecimento

Sempre que um novo elemento TRIZ derivado era identificado, ele era automaticamente incorporado ao repositório vetorial, promovendo a evolução contínua da base de conhecimento a cada patente processada.

Na Figura 27 é apresentada a representação esquemática da arquitetura híbrida.

Figura 27

Representação esquemática da arquitetura híbrida



Nota. Elaborado pela Autora (2025).

A Lógica de Decisão Derivativa implementa o mecanismo responsável por determinar se a representação TRIZ extraída pela LLM deve ser integrada ao repositório de conhecimentos ou substituída por um registro já existente no repositório semântico (*vector store*).

Os dados de entrada dessa lógica são:

(a) Uma extração estruturada proveniente da LLM (campos *task*, *object* e *physical_effect*);

(b) Um conjunto de amostras semelhantes recuperadas do *vector store*.

(c) A partir desse conjunto de amostras, o módulo decisor calcula uma medida de similaridade/distância. O *best_score* é definido como a menor distância entre *eq* e os *embeddings ei*, conforme Equação 6:

$$best\ score = \min dist\ (eq, ei), \text{ para i=1,2, ...N} \quad (6)$$

Onde: *eq* é o vetor de *embedding* da consulta (*query*) e {ei | i = 1, 2, ..., N} o conjunto de *embeddings* dos documentos recuperados. A métrica de distância empregada pelo *vector store* é a distância de cosseno, conforme a documentação oficial do LangChain Chroma *Vector store*[14] e representada na Equação 7:

$$dist\ (e_q, e_i) \ = \ 1 \ - \ cos\ \theta = 1 - \frac{e_q \cdot e_i}{|e_q||e_i|} \quad (7)$$

onde e_q é o vetor da consulta, e_i é o vetor recuperado e $\|e_q\|$ $\|e_i\|$ representa sua norma euclidiana.

Portanto, a fórmula final da distância de cosseno mínima, usada para determinar o *best_score* é definida na Equação 8:

$$\text{best\_score} = \min_i \left(1 - \frac{e_q \cdot e_i}{|e_q||e_i|}\right) \quad (8)$$

onde e_q é o vetor da consulta, e_i é o vetor recuperado e $\|e_q\|$ $\|e_i\|$ representa sua norma euclidiana.

(d) Para cada consulta, a função seleciona a amostra com menor distância (*best_doc*, *best_score*) e aplica uma regra de limiar (*threshold*), onde:

Se *best_score* < *threshold,* considera-se que existe um registro pré-existente suficientemente próximo. O sistema retorna os campos do registro mais similar e define *derived_from* = *null* (não gera nova derivação);

Caso contrário, considera-se que a extração da LLM representa uma nova relação semântica ou derivada. Nesse caso, a função retorna os campos extraídos pela LLM

---

14 https://api.python.langchain.com/en/latest/*vectorstores*/langchain_community.*vectorstores*.chroma.Chroma.html#langchain_community.*vectorstores*.chroma.Chroma.similarity_search_by_vector_with_relevance_scores

e registra em *derived_from* o conteúdo textual do documento mais próximo como referência de proveniência. O novo elemento é então inserido no *vector store*, ampliando a ontologia com o novo conhecimento.

Este fluxo de trabalho possibilita o mapeamento automatizado e escalável de patentes aos princípios da TRIZ, oferecendo suporte à busca e ao raciocínio semântico, além de garantir que a base ontológica evolua continuamente à medida que novas invenções são processadas.

O experimento demonstrou-se promissor, pois enriquece e atualiza os termos da ontologia com base em textos de patentes, assegurando que esses termos permaneçam alinhados à lógica e à terminologia próprias do domínio patentário, cujos resultados são apresentados na seção seguinte. A arquitetura do método, o código e a implementação estão disponibilizados em repositório público [https://github.com/PatrickLdA/patent-ai-project].

6.4.5 Avaliação e Análise de Desempenho do Método de Mineração Textual

Considerando os resultados promissores obtidos nos experimentos com a arquitetura LLM associada a um componente analítico, foi gerada uma amostra de documentos de patente para ser avaliada por especialistas, com o objetivo de validar se os relacionamentos semântico Tarefa e Objeto atribuídos a cada patente são compatíveis com o conteúdo do título e do resumo, e se o relacionamento ternário (T-O-EF), definido na ontologia, ou derivado, fornece uma sugestão de solução técnica que a patente possa endereçar. Adicionalmente, analisou-se se os termos T, O e EF adotados pelo componente analítico externo do LLM derivaram de registros da ontologia linguisticamente coerentes. Nas seções a seguir, será caracterizada a amostra de documentos de patente submetida à avaliação pelos especialistas, bem como apresentados os resultados alcançados.

6.4.5.1 Caracterização da amostra e distribuição por seção da Classificação Internacional de Patentes

A amostra analisada compreendeu 345 documentos de patente, correspondentes a 10,45% do total de documentos extraídos da base de patentes do INPI. Nessa amostra, o modelo de LLM com um componente analítico externo gerou 314 novos relacionamentos semânticos (T, O e EF) vinculados a um relacionamento original da ontologia e manteve 31 relacionamentos semânticos originais da ontologia.

Para analisar documentos de patente mais recentes, foram selecionados aleatoriamente documentos publicados em 2020, abrangendo todas as seções principais da IPC, que organiza as invenções conforme seu campo técnico e as divide em classes, subclasses, grupos principais

e subgrupos. A Tabela 19 apresenta a distribuição dos documentos de patente analisados por campo técnico.

Tabela 19

Distribuição dos documentos de patente analisados segundo as seções principais da Classificação Internacional de Patentes e principais classes

Seção	Seção IPC (1)	Quantidade de documentos	Percentual da amostra analisada (%)	Principais classes da IPC (2)
A	Necessidades humanas	92	26,7%	A61 - Ciência Médica ou Veterinária; Higiene A01 - Agricultura; Silvicultura; pecuária
B	Operações e transporte	62	18,0%	B01 - Processos Físicos ou Químicos em Geral B60 - Veículos Rodoviários ou Ferroviários
C	Química e metalurgia	72	21,0%	C07 - Química Orgânica C12 - Bioquímica; Cerveja; Bebidas Alcoólicas
D	Têxteis e papel	05	1,5%	D01 - Fios ou Filamentos; Seu Tratamento ou Processamento
E	Construções fixas	15	4,3%	E21 - Métodos ou Aparelhos para Perfurar a Terra ou Rocha
F	Engenharia mecânica, iluminação, aquecimento, armamento, explosivos	14	4,0%	F16 - Elementos e Comunicações de Máquinas F02 - Motores de combustão F03 - Máquinas ou motores para líquidos; motores movidos a vento, molas, pesos ou outros.
G	Física Computação	33	9,5%	G06 - Computador Digital; Processamento de Dados G01 - Medição; Teste
H	Eletricidade Telecomunicações	52	15,0%	H04 – Telecomunicações H01 - Componentes Elétricos
TOTAL		**345**	**100%**	

Nota. A Seção compreende o nível hierárquico mais alto na IPC, categorizando as invenções de acordo com o seu campo técnico (WIPO, 2025a).
A Classe compreende a segunda camada da IPC, que representa um domínio tecnológico dentro de uma seção, identificada por um código formado por uma letra e dois dígitos numéricos (WIPO, 2025a).

Os documentos de patente da amostra analisada evidenciam a dominância da Seção A (Ciências da Vida), na qual a classe A61 (Ciência Médica ou Veterinária; Higiene) se destaca, com 49 documentos, refletindo um foco significativo em tecnologias médicas, farmacêuticas e de saúde.

Como segunda classe mais representativa, observa-se A01 (Agricultura; Silvicultura; Pecuária), com 22 documentos.

A área de Tecnologia da Informação e Comunicação (Seção H) apresenta um volume expressivo, dividido principalmente entre as classes H04 (Telecomunicações), com 36 documentos, indicando um número substancial de inovações em redes, transmissão de dados e

comunicações, e H01 (Componentes Elétricos), com 11 documentos.

Por fim, a área de Química e Materiais (Seção C) reúne 35 documentos, com ênfase nas classes C07 (Química Orgânica) e C12 (Bioquímica; Cerveja; Bebidas Alcoólicas).

A Tabela 20 mapeia os domínios tecnológicos e detalha a composição integral do portfólio de patentes atribuído a cada especialista.

Tabela 20

Campos tecnológicos e concentração de documentos de patente por especialista

Código do especialista e formação	Foco principal	Caracterização dos documentos de patente selecionados e frequência
E1 (Engenharia mecânica)	B01 (Processos Físicos/Químicos) (n=10)	Foco em Engenharia e processos, com ênfase nas classes IPC B01 (n=10), A61 (n=5) e B65 (Transporte/Embalagem) (n=3).
E2 (Engenharia mecânica)	G06 (Computação) (n=5)	Ênfase em tecnologia e engenharia, com foco nas classes G06 (n=5), B60 (Veículos) (n=3), A01 (n=3) e B01 (n=3).
E3 (Engenharia elétrica)	H04 (Telecomunicações) (n=19)	Foco em Tecnologia da Informação e Comunicação, complementado por G06 (Computação) (6) e H01 (Componentes Elétricos) (7).
E4 (Farmácia)	A61 (Ciência médica ou veterinária) (n=12) e A01 (Agricultura) (n=9)	Distribuição contempla documentos de patentes da Ciências da Vida e Agronomia, e Química Orgânica. Também foram direcionados documentos da classe H04 (Telecomunicações) (n=5).
E5 (Farmácia)	A61 (Ciência médica ou veterinária) (18)	Foco em Ciências da Vida, com forte presença também em C07 (Química Orgânica) (n=6) e C12 (Bioquímica) (n=5).
E6 (Farmácia)	C07 (Química Orgânica) (n=7)	Foco em química e materiais, com ênfase em C07 (7) e C08 (Macromoléculas) (2). Inclui também documento de patente das classes B60 (n=3) e B29 (Plásticos) (n=2).
E7 (Química)	C07 (Química orgânica), A61 (Ciência médica ou veterinária), C11 (n=3 cada)	Distribuição diversidade em química e materiais, cobrindo várias classes (C01, C03, C05, C21, C22, C23, C25) com frequências baixas/médias (n=1 a 3).

Nota. Dados da pesquisa (2025).

A distribuição dos documentos de patente entre os especialistas foi prioritariamente orientada pela área de conhecimento de cada avaliador. Contudo, devido à natureza mais abrangente de alguns documentos e à distribuição não uniforme das patentes entre os campos técnicos, alguns documentos tiveram de ser alocados aos pesquisadores independentemente de sua formação específica para garantir o rateio equilibrado da carga de trabalho.

6.4.5.2 Condução do processo de avaliação por especialistas

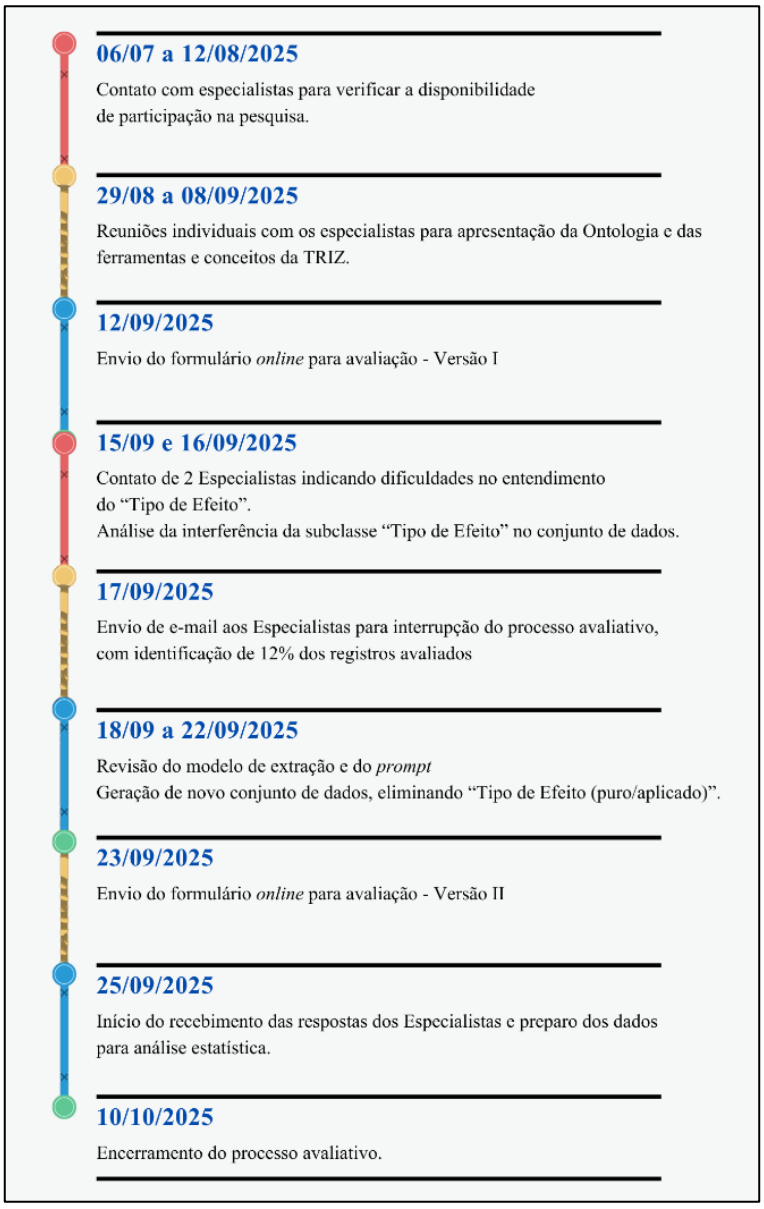
O processo de avaliação, contado a partir do envio da primeira versão do formulário online, transcorreu em aproximadamente 28 dias. Anteriormente, a partir de 6 de julho de 2025, foram realizados contatos com profissionais de nível superior familiarizados com patentes, totalizando 16 convites, dos quais sete confirmaram participação. As reuniões iniciais com os especialistas foram realizadas de forma presencial ou remota, em razão da dispersão geográfica dos avaliadores. Nessas reuniões, foi apresentada a ontologia, bem como os principais conceitos e ferramentas da TRIZ, até então desconhecida por cinco especialistas.

Logo após o envio do formulário, dois especialistas entraram em contato observando que o tipo de efeito — "puro" ou "aplicado" — adicionado à função de extração de termos do *prompt*, definida como Efeito físico = f (T, O), gerava ambiguidades e dificultava a análise. Imediatamente, foi realizada uma revisão da amostra de patentes, verificando-se que essa informação não era relevante no contexto da análise de compatibilidade entre os termos extraídos e o conteúdo textual do documento de patente. Assim, a informação foi removida, e um novo formulário online — denominado versão 2 — foi encaminhado aos especialistas.

Na Figura 28 é apresentado o desenvolvimento temporal do processo avaliativo, desde a elaboração do instrumento de avaliação até a coleta e análise das respostas dos especialistas.

Figura 28

Desenvolvimento temporal do processo avaliativo



Nota. Dados da pesquisa (2025).

Relatos dos especialistas E3, E6 e E7 indicaram dificuldades específicas na avaliação de documentos de patente alocados fora de seus campos tecnológicos de domínio. Reconhecendo que essa situação poderia comprometer a qualidade da avaliação, e visando assegurar a imparcialidade, foi sugerida a consulta à ontologia e ao campo de descrição do efeito físico como recursos de auxílio.

6.4.5.3 Análise estatística da consistência e da precisão do método de mineração textual

As respostas atribuídas às quatro questões do formulário de avaliação (Q1–Q4) foram organizadas em uma tabela, a fim de possibilitar análises estatísticas de desempenho do método

de mineração. Considerando que o formulário apresenta respostas binárias, e que seu objetivo é verificar se a extração dos componentes semânticos T, O e EF é consistente ou não consistente com o conteúdo do título e do resumo da patente, os dados obtidos permitem calcular a taxa de acerto da consistência e a taxa de erro dos termos extraídos, bem como a taxa de precisão (Miric et al., 2023) e a pontuação *Exact Match* (EM) (Lee & Bai, 2025).

Com base nesses dados, foi possível calcular a taxa de acerto da consistência, representada pelo VP, que corresponde ao termo identificado pelo método como positivo e classificado pelo especialista como consistente (escala 1), e a taxa de erro, associada ao FP, que abrange os termos definidos pelo método como consistentes, mas avaliados pelo especialista como inconsistentes (escala 2) ou inválidos (escalas 3 e 4), conforme aplicado por Miric et al. (2023). A Tabela 21 apresenta a distribuição das respostas dos especialistas por questão (Q1–Q4).

Tabela 21

Distribuição das respostas dos especialistas por critério de Verdadeiro Positivo e Falso Positivo

Questões / Respostas	(Q1) As subclasses Tarefa (T) e Objeto (O) atribuídas à patente são compatíveis com o título?	(Q2) As subclasses Tarefa (T) e Objeto (O) atribuídas à patente são compatíveis com o resumo?	(Q3) O relacionamento ternário atribuído à patente fornece uma sugestão de solução técnica que é efetivamente abordada ou resolvida pela própria patente?	(Q4) Existe consistência entre os termos derivados (T, O, EF) e aqueles dos quais se originam?
(1) Sim (VP)	179	286	293	115
(2) Não (FP)	42	32	52	199
(3) Informações do campo textual analisado são insuficientes (FP)	70	05	-----	----
(4) O relacionamento T–O é demasiado genérico (FP)	54	22	----	----
TOTAL	**345**	**345**	**345**	314 [(a)]

Nota. Dados da pesquisa (2025).
[(a)] Contabilizados apenas os relacionamentos gerados pelo componente analítico externo do método LLM.

A Tabela 22 apresenta os valores de Verdadeiros Positivos (VP) e Falsos Positivos (FP) obtidos nas quatro questões (Q1–Q4) do formulário de avaliação, bem como o total de documentos analisados, a taxa de precisão calculada para cada questão e a taxa de EM.

Tabela 22

Cálculo da Precisão e do *Exact Match* por questão Q1-Q4

Métrica	Q1	Q2	Q3	Q4	Interpretação
Verdadeiros Positivos (VP)	179	286	293	115	Número de documentos corretamente classificados como consistentes ("Sim").

Métrica	Q1	Q2	Q3	Q4	Interpretação
Falsos Positivos (FP)	166	59	52	199	Número de documentos incorretamente classificados como consistentes ("Sim", mas avaliado pelo especialista como inconsistente "Não" ou inválido (escalas 3 e 4).
Precisão (%)	51,97%	82,9%	84,93%	36,62%	Qualidade dos resultados positivos: proporção de documentos corretamente classificados como consistentes em relação ao total de documentos classificados como consistentes pelo método.
EM (decimal)	0,5197	0,8290	0,8493	0,3662	Taxa de documentos com classificação positiva correta: proporção de documentos classificados corretamente como consistentes em relação ao total de documentos avaliados.
Total Avaliado (VP + FP)	**345**	**345**	**345**	**314**	Base de documentos classificados como Positivos pelo método de mineração textual.

Nota. Dados da pesquisa (2025).

Os dados apresentados na Tabela 22 indicam que os avaliadores demonstraram um alto grau de concordância (acima de 82,9%) nos critérios que dependem da qualidade do resumo (Q2) e da funcionalidade da ontologia em sugerir uma solução técnica (Q3). Em Q2, observa-se boa correspondência entre os termos extraídos automaticamente e as classificações atribuídas pelos especialistas. Em Q3, o baixo número de falsos positivos (52) indica que o relacionamento ternário extraído (T-O-EF) é consistente, especialmente quando comparado ao conteúdo do resumo.

A compatibilidade com o título (Q1) é marginal (51,97%). A elevada quantidade de falsos positivos (166) em Q1, dos quais 124 registros derivam de critérios de avaliação relacionados a respostas ambíguas e genéricas (“Informações insuficientes” e “Relacionamento demasiado genérico”, conforme Tabela 21, sugere que a brevidade do título é provavelmente uma fonte significativa de incerteza na avaliação das subclasses T e O.

Na análise da consistência entre o relacionamento semântico ternário (T-O-EF) atribuído pelo método de mineração textual e aqueles derivados da ontologia, observa-se baixa precisão (36,62%). Em Q4, as 199 respostas “Não” evidenciam a principal fragilidade do método, sugerindo que os novos termos T, O e EF gerados pelo processo resultam em inconsistências com suas origens.

No contexto avaliado, cada predição foi analisada de forma binária (correto/incorreto), de modo que a taxa de precisão e a pontuação EM coincidiram, ainda que não sejam métricas equivalentes. Para o cálculo da EM, considerou-se que os valores de VP representam as predições “exatamente corretas”, visto que o formulário de avaliação prevê apenas uma resposta possível (1 – consistente; 2, 3 e 4 – não consistente).

Com base nos dados apresentados na Tabela 21, a precisão global do método foi calculada a partir da soma de todos os VP e FP das questões Q1 a Q3, conforme definido na

Equação 9:

$$\text{Precisão global} = \frac{\sum \text{VP}}{\sum(\text{VP} + \text{FP})} \quad (9)$$

A precisão global do método é de 73,26%. Como, neste contexto, a pontuação EM corresponde ao mesmo cálculo da precisão, o EM global médio, calculado conforme definido na Equação 10, é de 73,26%:

$$\text{EM global} = \frac{\sum(\text{EM x total})}{\sum \text{total}} \quad (10)$$

Ambas as métricas (Precisão e EM) indicam que o método apresenta, em média, consistência em aproximadamente dois terços das patentes analisadas.

Essas observações motivaram uma análise mais detalhada da consistência dos relacionamentos semânticos produzidos pelo método. Para verificar a consistência do relacionamento semântico ternário gerado pelo processo de mineração, foram considerados 314 registros (equivalentes a 91% da amostra analisada) nos quais o método não adotou as subclasses e o relacionamento original da ontologia, mas gerou um novo relacionamento semântico extraído do corpo textual e associou a um relacionamento existente na Ontologia. A Tabela 23 apresenta a precisão dos três critérios de avaliação (Q1, Q2 e Q3) do método de mineração textual.

Tabela 23

Cálculo da Precisão e do *Exact Match* do método de mineração com atribuição de novas subclasses Tarefa e Objeto e novo relacionamento semântico Tarefa–Objeto–Efeito Físico

Métrica	Q1	Q2	Q3	Interpretação
Verdadeiros Positivos (VP)	174	269	274	Número de documentos corretamente classificados como consistentes ("Sim").
Falsos Positivos (FP)	140	45	40	Número de documentos incorretamente classificados como consistentes ("Sim", mas avaliado pelo especialista como inconsistente "Não" ou inválido (escalas 3 e 4).
Precisão (%)	55,41%	85,67%	87,26%	Qualidade dos resultados positivos: proporção de documentos corretamente classificados como consistentes em relação ao total de documentos classificados como consistentes pelo método.
EM	0, 5541	0,8567	0,8726	Taxa de documentos com classificação positiva correta: proporção de documentos classificados corretamente como consistentes em relação ao total de documentos avaliados.
Total Avaliado (VP + FP)	**314**	**314**	**314**	Base de documentos classificados como Positivos pelo método de mineração textual.

Nota. Dados da pesquisa (2025).

As questões Q2 e Q3 apresentaram alta precisão, indicando que as etapas que dependem de informações textuais mais ricas, como o resumo ou o relacionamento ternário (T–O–EF),

são as mais confiáveis do processo.

A questão Q1, por outro lado, apresentou a menor precisão do conjunto (55,41%). Com 140 FP, essa etapa representa a principal fonte de erro no processo de extração. O elevado número de FP sugere que a extração baseada apenas no título é inadequada para capturar relações semânticas complexas, sendo inerentemente mais ambígua e menos confiável, o que reforça a necessidade de ajustes para reduzir a taxa de falsos positivos.

A questão Q2 manteve alta precisão, evidenciando excelente correspondência entre os termos extraídos pelo método e as classificações realizadas pelos especialistas, confirmando que o resumo constitui uma fonte de dados de alta qualidade para a tarefa de extração de termos.

Com 87,26% de precisão e apenas 40 FPs, a questão Q3 apresentou o melhor desempenho geral, validando que o relacionamento ternário extraído do conteúdo textual analisado (T–O–EF) fornece uma sugestão de solução técnica que é abordada no resumo da patente analisada.

Para avaliar a consistência do relacionamento ternário original da Ontologia, foram analisados apenas os registros em que o método adotou o relacionamento semântico originalmente definido, totalizando 31 registros (aproximadamente 9% da amostra analisada). A Tabela 24 apresenta a avaliação da precisão para as três questões (Q1, Q2 e Q3).

Tabela 24

Cálculo da Precisão e do *Exact Match* do método de mineração com atribuição de uma subclasse Tarefa e Objeto e um relacionamento semântico Tarefa–Objeto–Efeito Físico da ontologia

Métrica	Q1	Q2	Q3	Interpretação
Verdadeiros Positivos (VP)	5	17	19	Número de documentos corretamente classificados como consistentes ("Sim").
Falsos Positivos (FP)	26	14	12	Número de documentos incorretamente classificados como consistentes ("Sim", mas avaliado pelo especialista como inconsistente "Não" ou inválido (escalas 3 e 4).
Precisão (%)	16,13%	54,84%	61,29%	Qualidade dos resultados positivos: proporção de documentos corretamente classificados como consistentes em relação ao total de documentos classificados como consistentes pelo método.
EM	0,1613	0,5484	0,6129	Taxa de documentos com classificação positiva correta: proporção de documentos classificados corretamente como consistentes em relação ao total de documentos avaliados.
Total Avaliado (VP + FP)	**31**	**31**	**31**	Base de documentos classificados como Positivos pelo método de mineração textual

Nota. Dados da pesquisa (2025).

A Tabela 24 evidencia o melhor desempenho na questão Q3, que avalia se o relacionamento ternário (T–O–EF) atribuído à patente fornece uma sugestão de solução técnica compatível com o conteúdo da própria patente. No entanto, considerando que apenas 9% dos

documentos de patente da amostra integram essa avaliação, o índice permanece baixo. Isso indica que, quando o método mantém os termos das subclasses T e O e o relacionamento ternário (T–O–EF) da ontologia, o modelo apresenta acerto em mais de 60% dos casos.

Na questão Q2, que analisa se as sublcasses T e O atribuídas pelo modelo (a partir da Ontologia) são consistentes com o conteúdo textual do resumo, a precisão alcançada foi de 54,84%. Esse resultado indica que o método classifica corretamente como consistentes (VP = 17) mais documentos do que erra (FP = 14), representando um desempenho razoável, embora ainda sujeito a uma taxa de FP significativa.

A questão Q1 apresentou a menor precisão do conjunto (16,13%). Nessa questão, é avaliado se as subclasses T e O, extraídas da ontologia, são consistentes com o conteúdo textual do título. Entre os FP observados, 16 foram classificados como inválidos (notas 3 ou 4 na escala de avaliação) e 10 como inconsistentes (escala 2), indicando uma fragilidade estrutural na correspondência semântica entre os termos ontológicos e o conteúdo reduzido e ambíguo dos títulos de patente.

A Tabela 25 apresenta a precisão nos três cenários avaliados: (i) precisão do método por questão Q1-Q4, (ii) precisão com a atribuição novas subclasses T e O e novo relacionamento semântico T–O–EF, e (iii) precisão considerando as subclasses e relacionamento semântico T–O–EF original da Ontologia.

Tabela 25

Cálculo da Precisão do método de mineração nas questões Q1-Q4, por cenário de relacionamento

Precisão / Questão	Precisão (345 registros)	Precisão com relacionamento semântico T-O-EF original da Ontologia (31 registros)	Precisão com atribuição de um novo relacionamento semântico T-O-EF (314 registros)
Q1 (Compatibilidade com título)	51,97%	16,13%	55,41%
Q2 (compatibilidade com resumo)	82,9%	54,84%	85,67%
Q3 (Funcionalidade ternária)	84,93%	61,29%	87,26%
Q4 (Consistência da derivação)	36,62%	---	---

Nota. Dados da pesquisa (2025).

Na análise de compatibilidade com o título (Q1), a precisão é mais baixa no cenário com o relacionamento semântico original (16,12%). O cenário com a atribuição de um novo relacionamento semântico T–O–EF (55,4%) apresenta um aumento significativo em relação à precisão obtida com o relacionamento original e supera ligeiramente a precisão global (51,9%).

Na análise de compatibilidade com o resumo (Q2), observa-se precisão global de 82,9%.

O uso do relacionamento semântico original reduz a precisão para 54,8%, enquanto a atribuição de um novo relacionamento semântico T–O–EF resulta na maior precisão (85,7%), superando tanto a precisão global quanto a obtida com o relacionamento original.

Na análise de funcionalidade ternária (Q3), a precisão global é de 85,0%. O relacionamento semântico original resulta em uma precisão inferior (61,3%), enquanto o novo relacionamento semântico T–O–EF eleva a precisão para 87,3%, representando o melhor desempenho para essa questão.

Em Q1, Q2 e Q3, a atribuição de um novo termo e um novo relacionamento semântico T–O–EF apresentou o melhor desempenho. Esses resultados sugerem que a redefinição ou a atribuição de novos relacionamentos semânticos (T–O–EF) na ontologia foi eficaz para aprimorar a precisão da avaliação nas questões relacionadas à compatibilidade (Q1 e Q2) e à funcionalidade (Q3), em comparação tanto com a precisão global quanto com o uso do relacionamento semântico original.

Considerando que a precisão e o EM apresentam resultados equivalentes nos cálculos individuais e globais, em razão de dependerem exclusivamente dos valores de VP, FP e do total avaliado, a Tabela 26 apresenta o resumo dos resultados finais. Para o cálculo das métricas, foram consideradas as respostas às questões Q1 a Q3, que avaliam especificamente a consistência dos termos atribuídos em relação ao texto de patente analisado.

Tabela 26

Comparativo de Precisão e *Exact Match* global e por cenário de relacionamento semântico

Métrica	Global (345 registros)	Relacionamento semântico T-O-EF original da Ontologia (31 registros)	Atribuição de um novo relacionamento semântico T-O-EF (314 registros)
Precisão	73,26%	44,09%	76,11%
EM	73,26%	44,09%	76,11%

Nota. Dados da pesquisa (2025).

A Tabela 26 indica que o método atinge sua menor precisão/EM (44,09%) ao empregar a definição original da ontologia. Este resultado sugere que a interpretação, a granularidade ou a rigidez da ontologia inicial pode estar contribuindo para um elevado número de FP ou para um desalinhamento do método com esse padrão. Em contraste, a atribuição de um novo relacionamento semântico T-O-EF resulta no melhor desempenho, alcançando 76,11% de Precisão/EM. O desempenho global do método, que é de 73,26%, é significativamente impactado pela baixa performance observada com a ontologia original, o que ressalta a

necessidade de ajustes na modelagem semântica.

Para determinar a distribuição percentual dos documentos de patente analisados em cada campo tecnológico, correspondentes às seções A, B, C, D, E, F, G e H da IPC, bem como avaliar o comportamento desses campos na manutenção do relacionamento semântico original da ontologia ou na geração de novos relacionamentos, a Tabela 27 apresenta a distribuição dos registros por área tecnológica e por cenário de relacionamento.

Tabela 27

Distribuição e proporção da frequência de atribuição de relacionamentos semânticos originais e de novos relacionamentos semânticos, por seção do IPC

Seção IPC	Número total de documentos e % da amostra	% Manutenção do relacionamento semântico original	% Atribuição de novo relacionamento semântico
A Necessidades humanas	92 (27%)	4,35% (4/92)	95,65% (88/92)
B Operações e transporte	62 (18%)	16,13% (10/62)	83,87% (52/62)
C Química e metalurgia	72 (21%)	15,28% (11/72)	84,72% (61/72)
D Têxteis e papel	5 (1%)	0% (0/5)	100% (5/5)
E Construções fixas	15 (4%)	6,67% (1/15)	93,33% (14/15)
F Engenharia mecânica, iluminação, aquecimento, armamento, explosivos	14 (4%)	14,29% (2/14)	85,71% (12/14)
G Física, Computação	33 (10%)	6,06% (2/33)	93,94% (31/33)
H Eletricidade, Telecomunicações	52 (15%)	1,92% (1/52)	98,08% (51/52)
Total de documentos	**345**	**31**	**314**

Nota. Dados da pesquisa (2025).

Com base nos dados apresentados na Tabela 27, infere-se que a generalidade e a falta de atualização dos termos da TRIZ podem ter impactado diretamente a capacidade da ontologia em representar adequadamente os relacionamentos técnicos extraídos das patentes. Observa-se que a atribuição de novos relacionamentos semânticos (T–O–EF) ocorreu em 91% dos documentos analisados (314 de 345), enquanto apenas 9% mantiveram o relacionamento original da ontologia. Esse resultado sugere que os conceitos originais da TRIZ, muitos dos quais formulados em meados do século XX, não se mostraram semanticamente adequados para descrever as tecnologias contemporâneas expressas nos textos de patente.

Além disso, conforme evidenciado nos resultados da Tabela 22, 58% dos termos originais da ontologia foram classificados como inconsistentes em relação ao conteúdo textual

do título e do resumo, o que reforça a existência de um desalinhamento semântico e terminológico entre a TRIZ clássica e o vocabulário técnico atual utilizado nos documentos de patente.

Na Tabela 28 são apresentados os valores de VP e de FP obtidos nas três questões (Q1–Q3) do formulário de avaliação, bem como o percentual de precisão e o EM calculado para cada seção do IPC.

Tabela 28

Cálculo da Precisão e do *Exact Match* do método de mineração considerando Q1-Q3, por seção do Classificação Internacional de Patentes

Seção IPC	Q1 Compatibilidade de T e O com o título			Q2 Compatibilidade de T e O com o resumo			Q3 Avaliação da funcionalidade ternária (T–O–EF)		
	VP	FP	Precisão e EM	VP	FP	Precisão e EM	VP	FP	Precisão e EM
A	60	32	65%	75	17	82%	77	15	84%
B	29	33	47%	50	12	81%	49	13	79%
C	37	35	51%	58	14	81%	59	13	82%
D	2	3	40%	5	0	100%	5	0	100%
E	6	9	40%	14	1	93%	15	0	100%
F	4	10	29%	8	6	57%	11	3	79%
G	18	15	55%	31	2	94%	32	1	97%
H	23	29	44%	45	7	87%	45	7	87%
Total Avaliado (VP + FP)	**345**			**345**			**345**		

No contexto de obtenção dos dados (VP, FP e o cálculo da Precisão), a métrica EM é calculada da mesma forma que a precisão. Assim, a Tabela 28, que consolida a precisão e o EM do método de mineração por campo tecnológico, por seção do IPC e por questão (Q1-Q3), revela variações significativas. Na questão Q1 (compatibilidade com o título), o desempenho foi de fraco a moderado (29% a 65%), indicando a ambiguidade semântica presente nos títulos. Em contraste, os índices de Q2 (compatibilidade com o resumo) demonstram melhor desempenho, em razão da maior riqueza informacional contida nos resumos. Na Q3 (avaliação ternária T–O–EF), o método apresentou sua maior robustez, variando entre 79% e 100% de precisão, o que confirma sua consistência na sugestão de soluções técnicas relevantes.

As seções D e E da IPC se destacaram: a seção D foi a única a alcançar 100% de precisão nas questões Q2 e Q3, seguida pela seção E, com 93% em Q2 e 100% em Q3. Comparando esses valores com os da Tabela 27, verifica-se que, na seção D, a totalidade dos relacionamentos semânticos ternários foi adaptada semanticamente, sem aplicação dos relacionamentos

originais da ontologia; já na seção E, apenas um relacionamento semântico original foi mantido, em contraste com a criação de outros 14 novos relacionamentos semânticos.

Para aprofundar a análise sobre a consistência dos relacionamentos semânticos originais da ontologia e dos novos relacionamentos gerados, foi calculada a precisão para ambos os cenários. A Tabela 29 apresenta os resultados do cálculo da precisão do método de mineração textual quando utiliza as subclasses T e O da ontologia e o relacionamento semântico ternário (T–O–EF) original para extrair soluções genéricas dos campos textuais de título e resumo das patentes analisadas.

Tabela 29

Cálculo da Precisão e do *Exact Match* do método de mineração com atribuição de subclasses Tarefa e Objeto e relacionamento semântico Tarefa-Objeto-Efeito Físico da ontologia, por seção da Classificação Internacional de Patentes

Seção IPC	Q1 Compatibilidade de T e O com o título			Q2 Compatibilidade de T e O com o resumo			Q3 Avaliação da funcionalidade ternária (T–O–EF)		
	VP	FP	Precisão e EM	VP	FP	Precisão e EM	VP	FP	Precisão e EM
A	2	2	50,00%	3	1	75,00%	2	2	50,00%
B	0	10	0,00%	4	6	40,00%	5	5	50,00%
C	3	8	27,27%	6	5	54,55%	7	4	63,64%
D	0	0	0,00%	0	0	0,00%	0	0	0,00%
E	0	1	0,00%	1	0	100,00%	1	0	100,00%
F	0	2	0,00%	0	2	0,00%	0	2	0,00%
G	0	2	0,00%	2	0	100,00%	2	0	100,00%
H	0	1	0,00%	1	0	100,00%	0	1	0,00%
Total Avaliado (VP + FP)	**31**			**31**			**31**		

Nota. Dados da pesquisa (2025).

O método de relacionamento semântico T–O–EF demonstra uma capacidade significativamente superior de extração de subclasses (T–O) quando utiliza o resumo, em comparação com o título das patentes, conforme observado nas análises anteriores. Nas seções E, G e H, as subclasses da ontologia atribuídas pelo método mostraram-se consistentes com o conteúdo textual do resumo (Q2), e o relacionamento semântico ternário (T–O–EF) da ontologia (Q3) foi considerado consistente pelos especialistas para fornecer uma sugestão de solução técnica.

A Tabela 30 apresenta os resultados do cálculo da precisão do método de mineração textual quando atribui novas subclasses T e O e novo relacionamento semântico ternário (T–O–EF) para extrair soluções genéricas dos campos textuais de título e resumo das patentes

analisadas.

Tabela 30

Cálculo da precisão e do *Exact Match* do método de mineração com atribuição de novas subclasses Tarefa e Objeto e novo relacionamento semântico Tarefa–Objeto–Efeito Físico, por seção da Classificação Internacional de Patentes

Seção IPC	Q1 Compatibilidade de T e O com o título			Q2 Compatibilidade de T e O com o resumo			Q3 Avaliação da funcionalidade ternária (T–O–EF)		
	VP	FP	Precisão e EM	VP	FP	Precisão e EM	VP	FP	Precisão e EM
A	58	30	65,91%	72	16	81,82%	75	13	85,23%
B	29	23	55,77%	46	6	88,46%	44	8	84,62%
C	34	27	55,74%	52	9	85,25%	52	9	85,25%
D	2	3	40,00%	5	0	100,00%	5	0	100,00%
E	6	8	42,86%	13	1	92,86%	14	0	100,00%
F	4	8	33,33%	8	4	66,67%	11	1	91,67%
G	18	13	58,06%	29	2	93,55%	30	1	96,77%
H	23	28	45,10%	44	7	86,27%	45	6	88,24%
Total Avaliado (VP + FP)	**314**			**314**			**314**		

Nota. Dados da pesquisa (2025).

Os resultados apresentados na Tabela 30 indicam que, em todos os campos tecnológicos, o método demonstra uma capacidade significativamente superior de extração de subclasses T e O quando utiliza o resumo, em comparação com o título das patentes, mantendo os índices observados em todas as análises anteriores.

Na validação das subclasses e do relacionamento semântico ternário atribuídos pelo método, com base nos termos extraídos do corpo textual analisado, a concordância entre os especialistas apresenta valores elevados na análise do conteúdo textual dos resumos (Q2). Em Q3, o desempenho observado (entre aproximadamente 84% e 100%) evidencia o potencial da arquitetura híbrida em extrair termos do corpo textual das patentes que fornecem sugestões de soluções técnicas efetivamente abordadas ou resolvidas por elas.

Em síntese, os dados validam a arquitetura híbrida como uma estratégia relevante para mineração de inteligência técnica em português, sendo mais robusta na análise de resumos e quando permite a derivação e adaptação de novos termos semânticos, o que sugere que os termos originais da ontologia, isoladamente, são limitados para a mineração.

A seguir, apresentam-se alguns exemplos de extração de relacionamentos semânticos obtidos por meio do método de mineração textual baseado na ontologia. A Tabela 31 apresenta dois exemplos em que o método de mineração textual atribuiu às subclasses T e O e o

relacionamento semântico ternário da ontologia.

Tabela 31

Casos ilustrativos de patentes com atribuição de relacionamentos semânticos da ontologia

Título	Resumo (trecho chave)	Termos da Ontologia	Avaliação do especialista
Método de produção de uma placa de resfriamento	Método para produção de uma placa de resfriamento (3) feita de um material com excelente condutividade térmica, como cobre, alumínio, sua liga ou semelhantes. A placa de resfriamento (3) tem pinos (4, 4 ', 4 ", 4'") que se projetam aproximadamente perpendicularmente sobre uma área de base em uma superfície ativa (5) revestida pelo meio refrigerante. Os pinos (4, 4 ', 4 ", 4'") são circundados por uma borda periférica plana (U) que se estende essencialmente radialmente e funcionalmente necessária	T - Produzir O – Sólido EF - Resfriamento	(Q1) – 4 (T e O genéricos em relação ao título) (Q2) – 4 (T e O genéricos em relaçãoao resumo) (Q3) – 1 (relacionamento ternário sugere uma solução técnica conceitual abordada pela patente)
Vórtice ascendente para um separador ciclônico e aspirador de pó	O vórtice ascendente compreende uma pluralidade de pás de hélice estacionárias (V) que têm uma extremidade frontal convexa redonda ao redor da qual o ar entrante (A) é guiado para o vórtice ascendente (F), sendo que, quando o ar se separa da pá de hélice (V) dentro do vórtice ascendente (F), uma seção transversal das pás de hélice (V) tem apenas uma borda afiada (E). De preferência, uma linha média (M) da seção transversal das pás de hélice (V) não cruza uma linha de corda (C) em uma metade a montante da seção transversal. De preferência, um lado das pás de hélice (V) voltado para o ar entrante (A) é dotado de uma protuberância (P) em um ponto de estagnação (S). A protuberância (P) pode ser conformada de modo a guiar o ar entrante (A) para dentro do vórtice ascendente (F), e pode ter	T – Separar O – Gás EF – Tubo de vórtice	(Q1) – 1 (T e O compatíveis com a matéria textual do título) (Q2) – 1 (T e O compatíveis com a matéria textual do resumo) (Q3) – 1 (relacionamento ternário sugere uma solução técnica conceitual abordada pela patente)

	um lado côncavo seguindo um formato de uma pá de hélice vizinha e um topo arredondado		

Nota. Dados da pesquisa (2025).

A Tabela 32 apresenta exemplos em que o método de mineração textual atribuiu novas subclasses T e O, além de um novo relacionamento semântico ternário, a partir da análise do campo textual (título e resumo), relacionando-os a um relacionamento existente na ontologia.

Tabela 32

Casos ilustrativos de patentes com geração de novos relacionamentos semânticos a partir do texto analisado

Título	Resumo (trecho chave)	Relacionamento semântico	Avaliação do especialista
Misturas de pesticidas, método para controlar fungos nocivos fitopatogênicos e método para proteger plantas em crescimento ou materiais de propagação de plantas do ataque ou infestação por pragas invertebradas	Misturas de pesticidas compreendendo como compostos ativos (...) presentes em uma proporção em peso de 1000:1 a 1:1000; métodos e uso dessas misturas para combater pragas invertebradas, tais como insetos, aracnídeos, nematoides e/ou fungos nocivos dentro e sobre as plantas, e para proteger tais plantas sendo infestadas com pragas e/ou fungos nocivos.	Novo relacionamento: T – Controlar O – Fungos fitopatogênicos EF - Inibição de processos metabólicos Original da Ontologia: T – Misturar O – Campo EF – Heteródino	(Q1) – 1 (T e O compatíveis com a matéria textual do título) (Q2) – 1 (T e O compatíveis com a matéria textual do título) (Q3) – 1 (relacionamento ternário sugere uma solução técnica conceitual abordada pela patente) (Q4) – 2 (inconsistência entre os termos derivados (T, O, EF) e aqueles dos quais se originam)
Órgão digital com polifonia total	compreendido por um órgão digital com polifonia total, é um projeto com os mais avançados recursos da área da programação de micro controladores, onde é utilizado equações embutidas para ativar os espaços vazios da memória RAM, possibilitando o posicionamento de teclas em multifunções, aumentando assim a polifonia.	T – Ativar O – Memória RAM EF - Aumento da capacidade de processamento Original da Ontologia: T – Dobrar, O – Campo EF - Processamento	(Q1) – 3 (informações do título são insuficientes) (Q2) – 1 (T e O compatíveis com a matéria textual do resumo) (Q3) – 1 (relacionamento ternário sugere uma solução técnica conceitual abordada pela patente)

		digital de imagem	(Q4) – 2 (inconsistência entre os termos derivados (T, O, EF) e aqueles dos quais se originam)
antígeno-anticorpo	A presente invenção refere-se a moléculas de ligação ao receptor órfão 1 (ROR1) de tipo tirosina quinase de receptor, em particular anticorpos humanos específicos para ROR1, incluindo fragmentos de anticorpos. A presente invenção refere-se ainda a receptores recombinantes, incluindo receptores antigênicos quiméricos (CARs), que contêm tais anticorpos ou fragmentos, e polinucleotídeos que codificam os anticorpos, fragmentos de ligação a antígeno ou receptores específicos para ROR1. A invenção refere-se ainda a células geneticamente modificadas que contêm tais proteínas e receptores de ligação à proteína ROR1 e métodos relacionados e usos dos mesmos em terapia celular adotiva.	Novo relacionamento: T - Reconhecer O – Proteínda ROR1EF – Ligação Original da Ontologia: T – Inibir, O – Receptor de adenosina, EF – Bloqueio de sinalização celular	(Q1) – 1 (T e O compatíveis com a matéria textual do título) (Q2) – 1 (T e O compatíveis com a matéria textual do resumo) (Q3) – 1 (relacionamento ternário sugere uma solução técnica conceitual abordada pela patente) (Q4) – 1 (consistência entre os termos derivados (T, O, EF) e aqueles dos quais se originam)

Nota. Dados da pesquisa (2025).

O conjunto completo de dados do processo avaliativo dos especialistas é disponibilizado em: [https://osf.io/hu6fk/overview?view_only=e2b5637f780845bdaf476f70730c9798].

6.4.6 Depoimentos dos Especialistas

A validação dos resultados por especialistas confirmou a consistência e relevância prática do método. Os relatos a seguir, obtidos a partir de depoimentos pessoais, ilustram o processo de avaliação e as percepções sobre a aplicação do método e da ontologia:

Especialista E1*:* A tarefa foi difícil. Num primeiro momento, fiquei confuso quanto aos termos extraídos, tive que retornar ao texto da patente diversas vezes, mas com o andamento da avaliação foi ficando claro que os termos eram representativos e davam uma ideia da solução técnica. A partir desta compreensão, tudo ficou mais fácil.

Especialista E2: Mudei minha lógica de organizar um relatório de patente quando conheci a Teoria TRIZ, que até então eu desconhecia. Gostei muito da organização da solução em subclasses, faz mais sentido até para construir a estratégia de busca. As palavras que identificavam a Tarefa eram fáceis de identificar, mas Objeto e Efeito físico nem tanto. Objeto estava bem genérico, e efeito físico tinha nomenclaturas pouco utilizadas.

Especialista E3: Grande parte dos documentos que analisei eram da área da computação, onde geralmente os textos são densos. Analisar somente pelo título e pelo resumo foi muito difícil, a princípio tudo era 3 ou 4. Chamei a Kátia por duas vezes, estudei um pouco mais a TRIZ, foi um desafio.

Especialista E4: Gostei da experiência. As patentes analisadas eram da minha área, em determinados momentos ficou fácil. Achei que às vezes faltavam opções de respostas, porque às vezes o termo "meio" que se aproximava.

Especialista E5: Foi uma das experiências mais interessantes que tive na área da PI. Nunca tinha pensado em patentes dessa forma, e acho que faz todo o sentido quando se pretende pesquisar efeitos físicos em milhões de documentos. A ontologia é desatualizada, os termos – principalmente de EF, eram desconhecidos e os termos de Objeto eram muito genéricos. Mas para começar uma mineração, achei bem interessante.

Especialista E6: As patentes analisadas misturavam assuntos da minha área e assuntos de outras áreas. Aí tive dificuldade. Tive de recorrer à ontologia muitas vezes e ler o descritivo do conceito do Efeito físico e às vezes procurar mais informações na Internet.

Especialista E7: Acho que fui a avaliadora mais complicada (rs). Comecei o processo de avaliação com muitas dificuldades e em todas as patentes tive dúvida na resposta mais adequada. Acionei a Kátia diversas vezes, mas acho que nas patentes que analisei, a ontologia não ajudou muito.

Os testemunhos coletados indicam que, apesar da dificuldade inicial na interpretação dos termos extraídos e da necessidade de recorrer à ontologia e à literatura sobre TRIZ (E1, E3, E6), o uso do modelo proporcionou novas formas de interpretar o conhecimento técnico (E5, E2), favorecendo a aprendizagem e a disseminação de boas práticas. As dificuldades apontadas, como a generalidade dos termos de "Objeto" e o desconhecimento de certas nomenclaturas de "Efeito Físico" na ontologia (E2, E5), e a necessidade de mais opções de resposta (E4), oferecem *feedback* para a iteração e expansão da base semântica. A arquitetura modular e o caráter aberto dos artefatos tecnológicos asseguram sua escalabilidade e replicabilidade, permitindo que as limitações práticas levantadas sejam superadas a longo prazo, com a incorporação do método a sistemas nacionais e internacionais de inteligência técnica.

6.5 Discussão

O estudo adotou um fluxo de trabalho progressivo, iniciando com técnicas clássicas de PLN e evoluindo para arquiteturas baseadas em LLMs, com o objetivo de extrair os componentes semânticos T, O e EF a partir de patentes em língua portuguesa, utilizando uma ontologia baseada nos efeitos físicos TRIZ como estrutura de referência.

A evolução metodológica ao longo do estudo evidenciou os desafios da mineração textual em documentos de patente redigidos em português, idioma que apresenta elevada variabilidade morfológica e complexidade sintática.

A primeira abordagem, baseada em marcação morfossintática com a Biblioteca *SpaCy*, mostrou-se ineficaz para textos técnicos em português, apresentando rotulação incorreta de classes gramaticais, o que comprometeu a identificação precisa dos componentes Tarefa (verbos) e Objeto (substantivos).

A segunda abordagem, o *Match Vetorial*, superou parcialmente as limitações lexicais, permitindo identificar candidatos a T e O por similaridade semântica (medida pelo cosseno). Entretanto, essa técnica ainda apresentou dificuldades na correta classificação dos papéis gramaticais, além de depender do vocabulário limitado dos modelos pré-treinados em português, resultando, em alguns casos, em vetores com norma nula.

A arquitetura híbrida, que integra um LLM (como o GPT-4) e um módulo de Lógica de Decisão Derivativa, demonstrou ser a solução mais robusta para a tarefa proposta. Essa combinação otimizou a capacidade interpretativa e de extração inicial do LLM com a consistência lógica e o controle do componente analítico externo. Como resultado, o sistema alcançou maior confiabilidade e permitiu uma expansão controlada da ontologia.

O método atinge uma Precisão global e um EM de 73,26%. Analisando o desempenho sob diferentes configurações:

(1) Quando a lógica de decisão derivativa extrai e atribui novos termos do texto de patente analisado (título e resumo), a Precisão/EM sobe para 76,11%.

(2) Nos registros que mantiveram os termos originais da ontologia, observou-se um decréscimo expressivo na Precisão/EM, com valor de 44,09%.

Esses resultados evidenciam que, ao permitir a atribuição de novas subclasses de T e O e novos relacionamentos T–O–EF, o método alcança ganhos significativos de desempenho em todos os quesitos avaliados.

A elevada taxa de criação de novos relacionamentos (aproximadamente 91%) demonstra que o método de mineração textual precisou realizar uma adaptação semântica substancial dos

conceitos da TRIZ para representar de forma mais adequada os conteúdos extraídos das patentes. Embora não seja possível concluir que essa adaptação decorra exclusivamente da generalidade ou da desatualização dos termos da ontologia, é importante considerar que o contexto tecnológico contemporâneo, marcado pela intensa produção científica e pelo aumento dos depósitos internacionais de patentes em áreas como tecnologia da computação, tecnologia médica, comunicação digital e tecnologias energéticas (World Intellectual Property Organization, 2024), provavelmente contribui para esse descompasso terminológico.

Para reforçar esta argumentação, a literatura recente destaca a necessidade de atualização conceitual das ferramentas TRIZ para acompanhar a evolução dos domínios tecnológicos emergentes (Berdyugina & Cavallucci, 2023; Trapp & Warschat, 2025).

Uma segunda questão a ser considerada refere-se à generalidade dos termos da ontologia, que, embora estejam relacionados ao domínio da ciência e da tecnologia, podem não oferecer uma estrutura suficientemente específica e sensível ao contexto, capaz de capturar os detalhes e os relacionamentos intrincados característicos de determinados campos tecnológicos (Kitamura et al., 2024; Taduri et al., 2019; Trappey et al., 2024; Trappey, Trappey, & Chang, 2020; Vincent & Cavallucci, 2018). De modo semelhante, na definição da subclasse “objeto”, genericamente representada por categorias como sólido, sólido dividido, área, líquido ou gasoso, faz-se necessária uma expansão terminológica mais aprofundada por meio de variações lexicais. Essa ampliação deve incluir a atribuição de hiperônimos, palavras de sentido mais abrangente que compartilham características comuns (por exemplo, ferramenta é hiperônimo de chave de fenda e alicate), e de hipônimos, termos de sentido mais específico, ligados por características próprias (por exemplo, flores e árvores são hipônimos de flora).

De modo geral, a arquitetura híbrida do método e o suporte ontológico mostrou-se eficaz para a identificação de padrões semânticos e para a inferência de relacionamentos técnicos consistentes, com precisão média de global de 73,26% (Tabela 26). Dentre os 314 novos relacionamentos ternários, 36,62% (Tabela 26) foram validados pelo especialista como VP (escala 1), ou seja, são consistentes com os termos da ontologia, indicando que a geração de novos termos semânticos nem sempre preservou a coerência com sua origem ontológica. Sobre tais aspectos podem ser imputadas dificuldades do modelo em distinguir adequadamente nuances semânticas ou relacionamentos mais abstratos entre os termos, o que pode ser testado com ajustes nos hiperparâmetros do modelo, tal como hiperparâmetros de arquitetura (p. ex.

número de camadas[15]), hiperparâmetros de geração (p. ex. temperatura[16]), entre outros.

Sobre os campos textuais analisados, a validação do método por especialistas revelou alto desempenho nos critérios dependentes de informações textuais mais ricas (resumo) e, contrariamente, o título apresentando precisão marginal, o que está em consonância com a literatura, segundo a qual a brevidade e a ambiguidade dos títulos frequentemente resultam em falsos positivos, classificados pelos especialistas como "informações insuficientes" ou "relacionamento demasiado genérico" (Kim et al., 2019; Liu et al., 2023).

As análises de desempenho do método de mineração por campo técnico evidenciam que, em todas as seções da IPC, o método gerou um número maior de novos relacionamentos semânticos em comparação à atribuição dos relacionamentos originais da ontologia, com destaque para as seções D (Têxteis e Papel), E (Construções Fixas), G (Física e Computação) e H (Eletricidade e Telecomunicações).

Esses resultados demonstram que o mecanismo de Lógica de Decisão Derivativa aprimorou significativamente a eficácia do método, permitindo a geração e validação de novos termos e a adaptação do modelo ao léxico contemporâneo e especializado das patentes em português. Embora a taxa de acerto da consistência dos novos relacionamentos semânticos ternários (T–O–EF) com os termos da ontologia tenha apresentado precisão relativamente baixa (36,62%, Tabela 25), o método de mineração textual, fundamentado em uma arquitetura híbrida, combinando LLM e lógica de decisão derivativa, e associado a uma ontologia, mostrou-se capaz de integrar interpretação semântica profunda, validação lógica estruturada e adaptação ontológica dinâmica, representando um avanço concreto na mineração de inteligência técnica em língua portuguesa. O conteúdo extraído pelo LLM, orientado por uma ontologia que integra dimensões lexicais e semânticas, atua como uma ponte entre a linguagem natural e o conhecimento técnico, permitindo identificar problemas e propor soluções potenciais em diversos campos tecnológicos. Dessa forma, o método supera as limitações de abordagens exclusivamente supervisionadas ou não supervisionadas e amplia o alcance interpretativo da IA, corroborando os achados de Trapp e Warschat (2025), Stamatis et al. (2024) e Miric et al. (2023).

[15] Quantidade de camadas de processamento do modelo.
[16] Controla a aleatoriedade da saída. Valores mais baixos (próximos de 0) tornam a saída mais determinística e focada no mais provável; valores mais altos aumentam a diversidade e criatividade.

6.6 Considerações Finais

O presente estudo alcançou seu objetivo de desenvolver e validar um método para a mineração de inteligência técnica a partir de documentos de patente redigidos em português. A transição metodológica, culminando na adoção de uma arquitetura híbrida (LLM + Módulo Analítico), mostrou-se a estratégia mais adequada para superar as barreiras linguísticas e a complexidade semântica inerentes aos textos técnicos.

A principal contribuição do estudo reside na validação da capacidade do método em extrair e gerar relacionamentos ternários (T–O–EF), os quais funcionam como sugestões de soluções conceituais alinhadas à TRIZ.

Apesar dos resultados promissores, o estudo identificou três áreas principais que requerem aprimoramento:

(a) Atualização da ontologia para preencher a lacuna terminológica e semântica existente para acompanhar a evolução dos domínios tecnológicos emergentes e para oferecer termos menos genéricos;

(b) Fragilidade na Derivação - a baixa precisão na questão de consistência da derivação (36,6%) indica a necessidade de refinamentos no componente analítico de decisão, de modo que os termos derivados (novos T, O e EF) preservem a coerência semântica em relação aos termos originais da ontologia.

(c) Ambiguidade do Título - a precisão limitada na análise baseada apenas no título reforça a insuficiência desse campo textual para sustentar uma extração semântica complexa, sobretudo devido à sua brevidade e natureza genérica.

Como direções para pesquisas futuras, recomenda-se:

(a) Aprimorar o componente de derivação, ajustando a Lógica de Decisão Derivativa para aumentar a precisão na validação de novos termos T, O e EF. O ajuste de hiperparâmetros é citado por Jiang et al. (2025).

(b) Explorar fontes textuais mais ricas, integrando o método a campos textuais mais extensos, como descrição técnica e reivindicações das patentes. Atualmente, o acesso restrito a esses campos na base de patentes do INPI/BR (disponíveis apenas em PDF) limita a análise; sua inclusão poderia fornecer o contexto necessário para mitigar as ambiguidades observadas em Q1.

(c) Estender a aplicação do método, considerando a robustez validada em Q2 e Q3, para a mineração e o enriquecimento contínuo da ontologia.

(d) Realizar a expansão terminológica da subclasse “objeto” da ontologia,

genericamente representada por categorias como sólido, sólido dividido, área, líquido ou gasoso, meio de variações lexicais, com a atribuição de hiperônimos e hipônimos.

Entre as principais limitações observadas, destaca-se o potencial viés no julgamento dos especialistas, decorrente de fatores como nível de formação, experiência prévia e até desatenção nas respostas, favorecida pelo anonimato do processo avaliativo.

Além disso, como em cada patente foi considerada a resposta de apenas um especialista, os resultados podem refletir viés individual. O ideal seria que cada documento fosse avaliado por, no mínimo, dois especialistas, permitindo verificar eventuais disparidades de interpretação. Contudo, a validação por especialistas enfrenta restrições práticas, como o número limitado de profissionais qualificados e o tempo elevado demandado pela análise.

Por fim, recomenda-se incorporar uma etapa de iteração no método, de modo que os resultados obtidos na primeira rodada de avaliação possam retroalimentar o modelo, permitindo ajustes e uma segunda rodada de validação mais refinada.

7 CONCLUSÕES E RECOMENDAÇÕES DA TESE

O conhecimento, enquanto recurso estratégico das organizações, é o elemento central da Visão Baseada no Conhecimento (KBV), que sustenta teoricamente esta Tese (Takeuchi, 2013; Zheng et al., 2011). Essa perspectiva concebe o conhecimento como um ativo dinâmico, intangível e socialmente construído, essencial para a criação de vantagem competitiva sustentável. A KBV enfatiza a geração de novos conhecimentos como fonte de inovação organizacional (Azmi et al., 2024; Cabrilo & Dahms, 2018), destacando que esse recurso pode ser adquirido, combinado e transformado em capacidades dinâmicas (Stoian et al., 2024). Nesse sentido, a criação, o compartilhamento e a aplicação do conhecimento configuram pilares fundamentais para a inovação e o desenvolvimento tecnológico.

A integração entre a KBV e o DSR fundamenta o propósito central desta tese: transformar o conhecimento técnico e tecnológico contido em patentes em um recurso acessível, aplicável e socialmente útil. Ao reconhecer o conhecimento como ativo estratégico e fonte de inovação, a pesquisa evidencia a necessidade de desenvolver artefatos capazes de ampliar o acesso e a utilização de informações complexas, tradicionalmente restritas a públicos altamente especializados. Assim, a democratização do conhecimento emerge como a classe de problemas abordada pelo DSR, orientando a concepção de soluções tecnológicas que atendem a demandas práticas e, simultaneamente, contribuem para o avanço teórico na gestão e mineração do conhecimento.

A inteligência técnica contida nas patentes, enquanto recurso informacional, pode ser convertida em conhecimento aplicável à criação de tecnologias inovadoras e ao aprimoramento de produtos e processos. O acesso ampliado a esse conhecimento beneficia não apenas a comunidade científica, mas também o setor produtivo e a sociedade, ao promover a integração e o intercâmbio de saberes. O crescimento expressivo no número de depósitos de patentes reforça, portanto, a necessidade de soluções tecnológicas capazes de extrair, interpretar e transformar grandes volumes de dados técnicos em conhecimento estratégico, subsidiando o planejamento organizacional e a formulação de políticas de inovação.

Sob a perspectiva da KBV, o conhecimento é o principal recurso estratégico das organizações; contudo, a literatura clássica sobre capacidade absortiva assume, de forma implícita, que o conhecimento externo está prontamente disponível e acessível. Os resultados desta tese demonstram que embora o conhecimento tecnológico esteja disponível, ele permanece estruturalmente inacessível em razão de barreiras linguísticas, cognitivas e técnicas inerentes aos documentos de patente.

Nesse cenário, o método e a ontologia desenvolvidos nesta pesquisa não “criam” capacidade absortiva, mas atuam como mecanismos habilitadores, ao remover barreiras antecedentes à aquisição. Ao transformar inteligência técnica de documentos patentários complexos em conhecimento estruturado, interpretável e reutilizável, o artefato amplia as condições necessárias para que as organizações reconheçam, assimilem, transformem e explorem o conhecimento externo de forma sistemática.

A solução proposta fundamenta-se em uma ontologia derivada dos efeitos físicos da TRIZ, que atua como núcleo conceitual e elemento orientador do artefato. Essa ontologia não apenas organiza o conhecimento técnico, mas também regula e qualifica a interpretação semântica realizada pela IA promovendo maior precisão na extração de informações e na geração de conhecimento aplicável.

No âmbito desta tese, foram criados e validados dois Produtos Técnico-Tecnológicos: (i) uma ontologia semântica estruturada a partir dos efeitos físicos da TRIZ; e (ii) um método para a mineração de inteligência técnica de patentes, com foco na extração do relacionamento semântico ternário Tarefa–Objeto–Efeito Físico (T–O–EF). Ambos foram concebidos e avaliados conforme as etapas do DSR propostas por Peffers et al. (2007), com validação por especialistas e praticantes, que confirmaram sua relevância prática e consistência metodológica.

Dessa forma, a pesquisa percorre integralmente as etapas do DSR, oferecendo simultaneamente uma solução prática e um avanço teórico no campo da mineração de textos de patentes, contribuindo para posicionar o Brasil no cenário internacional dessa área de pesquisa.

Para sintetizar os principais achados e contribuições, a Tabela 33 apresenta a Matriz Contributiva da tese, oferecendo uma visão integrada da estrutura do trabalho e demonstrando como os estudos foram articulados para responder à questão principal de pesquisa, conferindo caráter inédito e originalidade científica à investigação (Costa et al., 2024).

Tabela 33

Matriz Contributiva da Tese

QUESTÃO PRINCIPAL DE PESQUISA Como viabilizar a aquisição de inteligência técnica de patentes, de modo a permitir sua internalização e articulação com o conhecimento interno da organização, por meio de um método de mineração textual apoiado em uma ontologia?				
OBJETIVO PRINCIPAL Propor um método de mineração textual, apoiado em ontologia semântica e linguística, para viabilizar a aquisição e internalização das informações técnicas contidas em patentes no contexto organizacional..				
CONCLUSÕES ESPECÍFICAS				
Campo de pesquisa ou título	**Contribuições científicas**	**Contribuições técnicas, tecnológicas e/ou sociais**	**Limitações**	**Agenda futura**
ESTUDO 1 REVISÃO SISTEMÁTICA DA LITERATURA SOBRE MINERAÇÃO DE CAMPOS TEXTUAIS DE DOCUMENTOS DE PATENTE	• Consolida pesquisas sobre mineração de patentes. • Mapeia a evolução, identificando avanços, o estado atual e lacunas persistentes.	• Contribuições técnicas: Sintetiza estratégias e recomendações de metodologias validadas para orientar futuras agendas de pesquisa.	• Estratégia de busca (termos amplos/restritivos) e aplicação de critérios (inclusão/exclusão) podem ter introduzido vieses, limitando o escopo final.	• Investigar relações lexicais (sinônimos, polissemia). • Desenvolver redes semânticas e integrar bases terminológicas com repositórios dinâmicos. • Construir ontologias de domínio específicas, alinhadas lexical e semanticamente à terminologia de patentes.
ESTUDO 2 INTEGRAÇÃO ENTRE TRIZ E MINERAÇÃO DE TEXTOS DE PATENTES: AVANÇOS, DESAFIOS E TENDÊNCIAS NA EXTRAÇÃO DE INTELIGÊNCIA TÉCNICA	• Apresenta visão abrangente das metodologias e desafios da integração entre TRIZ/mineração. • Confirma a área como promissora, impulsionada pela aplicação de ML/DL em bases de patentes.	• Contribuições técnicas: Fornece base empírica/conceitual para o aprimoramento de sistemas automatizados de recuperação de informações técnicas em bancos de patentes multilíngues, utilizando conceitos TRIZ.	• Estratégia de busca (termos amplos/restritivos) e seleção de critérios podem ter introduzidos vieses. • Exclusão de publicações que não apresentavam modelos ou ferramentas TRIZ aplicáveis à mineração podem ter limitado a generalização dos resultados obtidos. • Restrição à base de dados *Scopus* pode ter introduzido viés de seleção.	• Otimizar o uso de ferramentas PLN para extração de informações inventivas. • Ampliar estudos sobre correlações semânticas (sinônimos, polissemia, etc.). • Promover a atualização terminológica das ferramentas TRIZ. • Comparar a qualidade das informações extraídas de diferentes seções das patentes. • Priorizar a adaptação de metodologias e ferramentas

				para análise de patentes multilíngues.
ESTUDO 3 DESENVOLVIMENTO DE UMA ONTOLOGIA BASEADA NA TEORIA DA SOLUÇÃO INVENTIVA DE PROBLEMAS (TRIZ)	• Desenvolvimento de uma ontologia semântica funcional, fundamentada nos efeitos físicos da TRIZ, pioneira em língua portuguesa, servindo como base semântica para a extração automática de soluções genéricas a partir de documentos de patente em língua portuguesa.	• Contribuições técnicas: desenvolvimento de uma ontologia semântica funcional baseada em efeitos físicos, servindo como base linguística para a mineração de patentes.	• Necessidade de expansão contínua (abrangência lexical e diversidade de fontes) para maior cobertura semântica. • A avaliação empírica em contextos multilíngues e interdisciplinares ainda precisa ser aprofundada.	• Integrar a ontologia com outras bases de conhecimento (tesauros técnicos, dicionários, classificadores IPC). • Enriquecer o modelo por meio da associação com ontologias específicas de domínio.
ESTUDO 4 DESENVOLVIMENTO DE UM MÉTODO DE MINERAÇÃO DE INTELIGÊNCIA TÉCNICA EM PATENTES UTILIZANDO UMA ONTOLOGIA BASEADA NOS EFEITOS FÍSICOS DA TRIZ	• Desenvolvimento e validação de método de mineração híbrido (LLM + Lógica Derivativa) apoiado em ontologia para extração lexical e semântica.	• O método integra interpretação semântica (LLM), validação lógica e adaptação ontológica dinâmica, superando limitações de abordagens singulares. • Fornece base para mineração de inteligência técnica em patentes, apoiando atividades de P,D&I.	• Identificação de lacuna terminológica e semântica na ontologia (abstração/generalidade), que requer atualização para domínios emergentes. • Potencial viés individual no julgamento (decorrente de formação/experiência) e limitação de ter um único especialista por patente, sem iteração.	• Aprimorar o Componente de Derivação e hiperparâmetros do modelo (LLM) para maior precisão e coerência dos novos termos T, O e EF. • Realizar expansão terminológica da subclasse "Objeto" (variações lexicais, hiperônimos/hipônimos). • Testar o método em campos textuais mais extensos das patentes. • Estender a aplicação do método para o enriquecimento contínuo e dinâmico da Ontologia TRIZ. • Submeter a avaliação por no mínimo dois especialistas e incorporar iteração (retroalimentação) no método.
CONCLUSÃO INTEGRADORA				
CONSOLIDAÇÃO DOS PRINCIPAIS RESULTADOS • O estudo consolida o conhecimento sobre mineração de patentes, mapeando a evolução, identificando lacunas persistentes e fornecendo uma visão abrangente das metodologias e dos desafios na integração entre TRIZ e mineração textual. • Desenvolvimento e validação de um método de mineração de arquitetura híbrida (LLM + Lógica Derivativa), que integra interpretação semântica profunda,				

validação lógica e adaptação ontológica dinâmica. • Confirma que a combinação de TRIZ com mineração de textos é uma área promissora, impulsionada pela crescente aplicação de IA para exploração eficiente das bases de patentes. • O método e a ontologia desenvolvida fornecem a base empírica/conceitual para o aprimoramento de sistemas automatizados de recuperação de informações técnicas em bancos de patentes, apoiando diretamente as atividades de P, D & I e a mineração em contextos multilíngues.
AVALIAÇÃO DA PRODUÇÃO CIENTÍFICA, TÉCNICA E TECNOLÓGICA
• O estudo propõe uma nova fronteira metodológica com o desenvolvimento de um método de mineração híbrido que integra a interpretação semântica profunda de um LLM (avançado modelo de IA) com uma validação lógica estruturada e adaptação ontológica dinâmica. Essa arquitetura é original por superar as limitações das abordagens de ML puramente supervisionadas ou não supervisionadas • A pesquisa é pioneira no contexto da língua portuguesa ao desenvolver uma ontologia semântica funcional (baseada em efeitos físicos da TRIZ) e ao criar uma base linguística para mineração de patentes. Essa adaptação de ferramentas conceituais complexas para um idioma com especificidades como o português representa um avanço inédito nos estudos de mineração textual. • A pesquisa articula a lógica conceitual da TRIZ com as técnicas de ponta de Inteligência Artificial (LLMs e validação lógica), fornecendo uma base para a mineração de inteligência técnica de patentes, que apoia diretamente as atividades de PD&I, com forte convergência nos eixos técnico e tecnológico. • Pela sua função utilitária, o estudo contribui de forma indireta e fundamental para o avanço tecnológico geral e a inovação, fatores essenciais para a inovação e o avanço tecnológico em geral.

Nota. Adaptado de Costa et al. (2024).

7.1 Impacto da Pesquisa na Sociedade

O conhecimento científico e tecnológico é um ativo essencial para impulsionar o desenvolvimento econômico e garantir a sustentabilidade organizacional. Dessa forma, a pesquisa assume uma relevância estratégica, alinhando-se tanto aos critérios de impacto estabelecidos pela CAPES quanto às categorias propostas por Wickert et al. (2021) para a área de gestão. Segundo essa abordagem, o impacto vai além da métrica puramente acadêmica, abrangendo cinco dimensões principais que promovem a transformação: Acadêmica (avanço da ciência), Prática/Organizacional (influência nas ações e no pensamento das organizações), Social (contribuição para a solução de grandes desafios globais), em Política Pública (apoio à formulação de políticas) e Educacional (inovação em ensino e currículos). A combinação intrínseca dessas formas de impacto permite que a pesquisa em gestão contribua de maneira efetiva para problemas sociais relevantes, indo além do contexto tradicional de negócios.

Os Produtos Técnicos e Tecnológicos (PTTs) desenvolvidos no estudo - um método de mineração textual integrando LLM com Lógica de Decisão Derivativa, e uma ontologia TRIZ semântica funcional - materializam o impacto da pesquisa ao transformarem o conhecimento em ferramentas aplicáveis. Esses PTTs, que promovem o uso inteligente de dados tecnológicos em múltiplos domínios e contextos linguísticos, estabelecem uma ponte direta entre o rigor científico e as necessidades do mundo real, avançando a ciência e produzindo ativos tangíveis que são essenciais para o desenvolvimento econômico e a sustentabilidade, conforme preconizado pelos critérios de impacto da CAPES e pela visão expandida de Wickert et al. (2021).

As seções subsequentes detalharão a contribuição específica da presente pesquisa para cada uma das dimensões de impacto previamente definidas, analisando os resultados alcançados em cada contexto.

7.1.1 Impacto Prático e Gerencial

O método de mineração de arquitetura híbrida, que integra LLMs e Lógica de Decisão Derivativa, possibilita a análise automatizada e escalável de documentos de patente, cuja densidade técnica e complexidade linguística tradicionalmente dificultam a interpretação manual. Esse avanço oferece a organizações, centros de P&D e gestores de tecnologia uma ferramenta que

amplia significativamente a capacidade analítica sobre bases patentárias, fortalecendo a base técnico-científica necessária à gestão da inovação, à propriedade intelectual e à formulação de políticas públicas de CT&I.

Considerando que a arquitetura, o código-fonte do método e a ontologia estão disponibilizados em repositórios públicos, qualquer organização pode replicar, adaptar e incorporar a solução conforme suas necessidades específicas. Empresas de base tecnológica, PMEs, instituições públicas e centros de pesquisa podem integrar o processo de mineração semântica aos seus fluxos de análise de patentes, utilizando o método para extrair automaticamente relações T–O–EF e converter grandes volumes de documentos técnicos em conhecimento acionável.

Essa capacidade permite aprimorar o uso estratégico das informações contidas em bases de patentes, reduzindo custos operacionais, acelerando processos de prospecção tecnológica e ampliando a capacidade de inovação. Por se tratar de uma solução aberta, modular e replicável, o método pode ser integrado a sistemas já existentes, adaptado a diferentes domínios tecnológicos ou utilizado como base para o desenvolvimento de novas ferramentas de inteligência técnica, contribuindo para o fortalecimento do ecossistema nacional de inovação.

Além disso, o método apresenta aplicações diretas nas quatro dimensões da capacidade absortiva. Na aquisição, automatiza o monitoramento de bases de patentes e a identificação de informações relevantes, reduzindo a dependência de equipes altamente especializadas. Na assimilação, a estrutura semântica T-O-EF facilita a compreensão de soluções técnicas existentes, reduzindo a sobrecarga cognitiva de engenheiros e pesquisadores. Na transformação, o conhecimento extraído pode ser recombinado com o conhecimento interno das organizações, ampliando o repertório tecnológico disponível. Por fim, na exploração, a inteligência técnica gerada subsidia decisões estratégicas relacionadas a investimentos, inovação e posicionamento tecnológico.

O impacto do método transcende a análise técnica de patentes, estendendo-se a áreas como vigilância tecnológica, inteligência competitiva, gestão do conhecimento, empreendedorismo e formulação de estratégias de inovação. O uso da ontologia derivada dos efeitos físicos da TRIZ como base semântica permite abstrair terminologias específicas de domínio, viabilizando a transferência de conhecimento entre diferentes setores tecnológicos.

7.1.2 Impacto Social

Sob a perspectiva social, a pesquisa contribui para democratizar o acesso ao conhecimento tecnológico, tradicionalmente restrito a países com maior domínio do inglês e infraestrutura tecnológica avançada. Ao desenvolver uma ontologia funcional adaptada à língua portuguesa, o estudo promove inclusão e autonomia científica e tecnológica para os países de língua portuguesa. Esse impacto se manifesta no fortalecimento das capacidades de pesquisa e inovação em contextos regionais e emergentes, permitindo que grupos e instituições locais acessem, interpretem e apliquem conhecimento técnico relevante sem depender de traduções ou intermediários especializados. O método também pode contribuir para ações de transferência de tecnologia, formação de redes cooperativas e valorização do conhecimento nacional, ampliando os benefícios da inovação para a sociedade em geral, alinhando-se aos ODS.

7.1.3 Impacto Político e Transparência

O impacto político é observado na aplicabilidade do método e da ontologia em português como instrumento de apoio à formulação de políticas públicas de CT&I. O método e a ontologia podem subsidiar agências governamentais e instituições de fomento na avaliação de tendências tecnológicas, identificação de lacunas em áreas estratégicas e planejamento de investimentos em inovação. Adicionalmente, o estudo propõe uma metodologia transparente e reprodutível, em conformidade com os princípios FAIR (*Findable, Accessible, Interoperable, Reusable*) (GO FAIR International Support & Coordination Office, 2025), o que contribui para políticas voltadas à abertura e interoperabilidade de dados científicos e tecnológicos. A disponibilização pública do código-fonte e da base ontológica reforça a transparência e promove a colaboração entre universidades, empresas e governos.

7.1.4 Impacto Acadêmico

O impacto acadêmico desta pesquisa manifesta-se, primeiramente, no desenvolvimento de uma ontologia semântica funcional e de um método híbrido de mineração textual, os quais podem ser incorporados como ferramentas pedagógicas e metodológicas em programas de pós-graduação. Esses PTTs favorecem o aprendizado ativo e interdisciplinar em temas como inteligência técnica, mineração de dados textuais e análise semântica, sendo aplicáveis ao ensino e à pesquisa nas áreas

de Gestão da Inovação, Ciência da Informação, Linguística Computacional e Engenharia do Conhecimento. Ademais, os artefatos desenvolvidos oferecem uma base metodológica robusta para novos projetos acadêmicos, fomentando investigações futuras na interface entre Inteligência Artificial e inovação tecnológica.

Do ponto de vista teórico, a principal contribuição da tese reside na ampliação KBV e dos estudos sobre capacidade absortiva. A pesquisa evidencia que a existência e a disponibilidade formal de informação tecnológica não asseguram, por si só, sua absorção pelas organizações. Ao identificar empiricamente barreiras linguísticas, cognitivas e técnicas como antecedentes negligenciados da capacidade absortiva, a tese demonstra que o conhecimento patentário pode existir sem ser efetivamente absorvível. Nesse sentido, a ontologia e o método híbrido, que integra LLMs e Lógica de Decisão Derivativa, operam como uma infraestrutura cognitiva e semântica que viabiliza a conversão de informação técnica codificada em conhecimento organizacional aplicável.

Assim, a pesquisa contribui para o avanço teórico ao explicitar o papel das tecnologias de mineração semântica e da modelagem ontológica como mecanismos intermediários entre a disponibilidade de informação e a geração de conhecimento, oferecendo novos caminhos analíticos para estudos futuros sobre gestão do conhecimento, inovação e inteligência técnica.

7.1.5 Originalidade Científica e Tecnológica

Do ponto de vista científico e tecnológico, a pesquisa inaugura uma nova fronteira metodológica ao propor uma arquitetura híbrida de mineração textual, combinando a interpretação semântica profunda de LLMs com uma validação lógica derivativa. Essa originalidade metodológica supera as limitações das abordagens puramente supervisionadas ou não supervisionadas, ampliando a confiabilidade e a interpretabilidade dos resultados. A ontologia em português constitui uma iniciativa pioneira, servindo como referência para futuras pesquisas em ontologias técnicas e mineração semântica multilíngue, sendo sua adaptação ao português um avanço inédito que permite a expansão da pesquisa tecnológica nesse idioma de baixo recurso computacional.

O método proposto alcançou Precisão e EM globais de 73,26%, demonstrando desempenho competitivo em relação às abordagens de última geração. Na atribuição de novos relacionamentos semânticos T–O–EF, o método obteve 76,11%, enquanto, ao empregar os termos originais da ontologia, o desempenho foi de 44,09%.

A seguir, apresenta-se uma comparação entre o método de mineração textual desenvolvido e os estudos mais recentes que aplicam tecnologias de IA, com ênfase em LLMs:

(a) Miric et al. (2023) atingiram 83% de precisão utilizando *embeddings* GloVe pré-treinados de 300 dimensões, com 4.000 patentes do USPTO — 3.200 para treinamento e 800 para validação.

(b) Lee e Bai (2025) obtiveram 60,93% de *Exact Match* com o PAI-NET, um agente de recuperação composto por um módulo de engenharia de *prompt* e um LLM, treinado com 100.000 patentes da República da Coreia (2020–2023) e testado com outras 20.000.

(c) Trapp e Warschat (2025) alcançaram 46% de precisão utilizando o GPT-4 para identificar contradições técnicas TRIZ em 3.200 patentes em inglês (metade contendo contradições anotadas por especialistas). O *LangChain* foi empregado para o gerenciamento das interações.

(d) Blume et al. (2024) obtiveram precisão superior a 60% em tarefas de distinção entre documentos relevantes e 99,61% em tarefas mais restritas, com o modelo *Max Chunk-Claim CCX*, baseado em transformadores e aplicado a textos completos do EPO e USPTO.

(e) Li, Yu, et al. (2023) desenvolveram um método de rede lexical combinado com *BERT*, alcançando 82,73% de precisão; ao integrar CNN e LSTM, observaram acréscimos de 2,19% e 2,25%, respectivamente, na classificação de textos de patentes em chinês.

Embora a Precisão seja uma métrica fortemente influenciada pela arquitetura do modelo, pela natureza dos dados e pelos procedimentos de pré-processamento e avaliação, o método proposto nesta tese demonstra desempenho equivalente aos modelos mais avançados baseados em IA. Destaca-se por sua robustez semântica, menor dependência de *corpora* rotulados e capacidade de processamento em língua portuguesa, um contexto ainda pouco explorado na literatura científica.

Além disso, o método desenvolvido apresenta avanços significativos em relação às abordagens anteriores, conforme descrito a seguir:

(a) Em comparação com estudos que utilizam técnicas clássicas de PLN, como LSA e método do Cosseno, apresentados por Korobkin et al. (2019) e Korobkin e Fomenkov (2018), o método supera as limitações de precisão (47%–60%) observadas na extração

de efeitos físicos, especialmente em textos técnicos em português.

(b) Avança em relação aos estudos de mineração textual com aplicação da TRIZ, como os de Russo e Montecchi (2011), Russo et al. (2012), Korobkin, Fomenkov e Kolesnikov (2018), Korobkin et al. (2019), Berdyugina e Cavallucci (2021, 2022b, 2023), (Guarino et al. (2021, 2022, 2024), Guarino, Samet, Nafi, et al. (2020), Trapp e Warschat (2025) e Du et al. (2025), que negligenciam o tratamento de patentes em português.

(c) Diferencia-se das abordagens SAO/PF e Palavra-Chave, como as propostas por Kim, Kim, et al. (2018) e Lee et al. (2019), que dependem fortemente de *corpora* rotulados e apresentam menor flexibilidade semântica.

(d) Amplia os modelos SAO/PF descritos por Choi, Park, et al. (2012), Dewulf (2011), Park, Kim, et al. (2013), Vicente-Gomila et al. (2017) e Yoon (2010), ao adotar a tríade T–O–EF) fundamentada nos princípios da TRIZ, oferecendo uma representação semântica mais precisa e próxima dos conceitos inventivos em patentes.

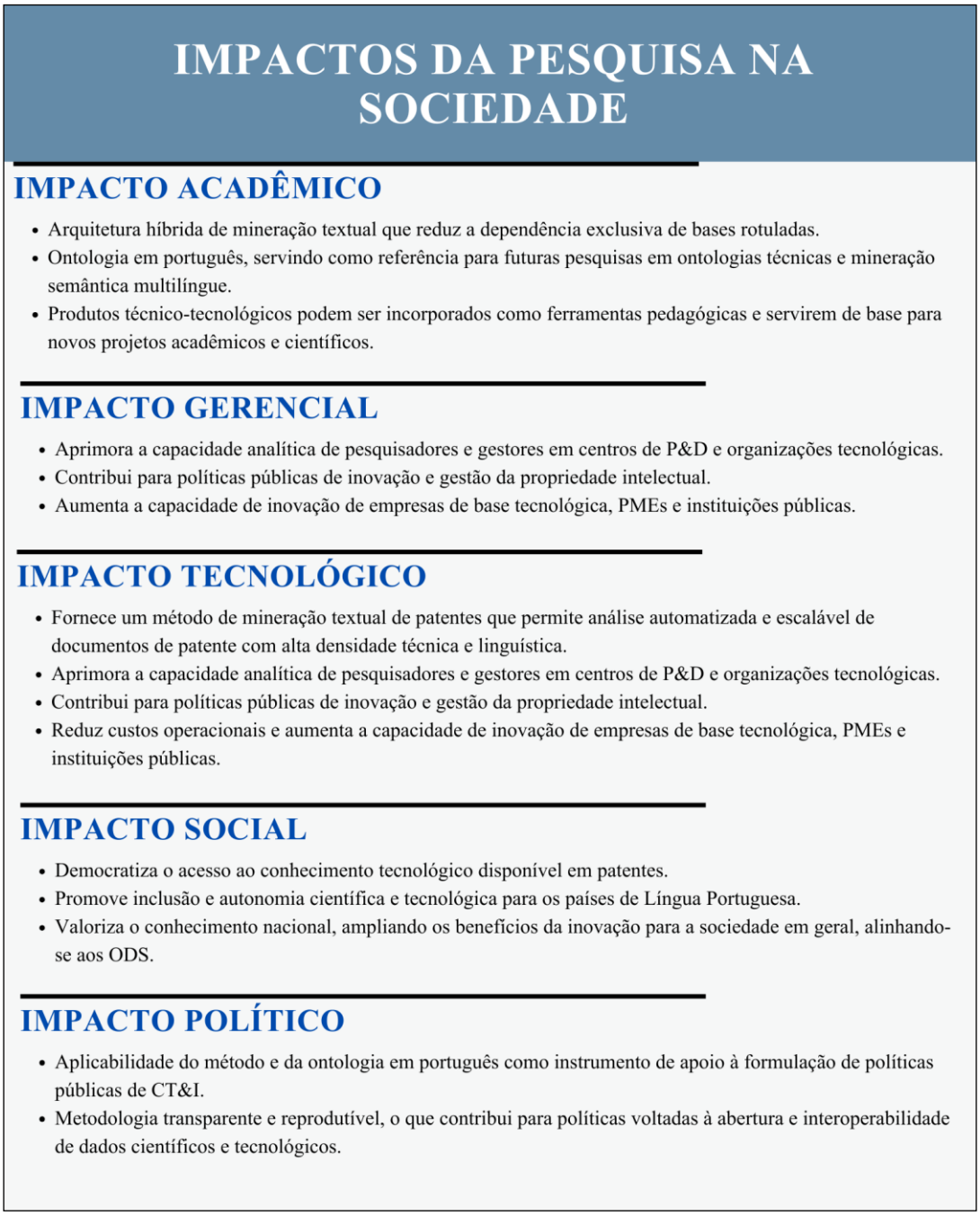
(e) Em relação às abordagens de DL, *embeddings* e redes neurais (Huang et al., 2023; Jiang et al., 2023; Jiang et al., 2025; Kaliteevskii et al., 2020, 2021; Nkolongo et al., 2024; Sarica et al., 2020), o método desenvolvido se destaca por reduzir a dependência de grandes bases de dados rotuladas, mantendo desempenho competitivo.

A tese demonstra um caminho alternativo: uma arquitetura híbrida que reduz a dependência de bases rotuladas, permitindo adaptação a diferentes domínios, mitigando a barreira linguística e ampliando a aplicabilidade.

Em síntese, esta pesquisa responde à questão central ao desenvolver e validar um método de mineração textual de inteligência técnica de patentes, apoiado em uma ontologia semântica fundamentada nos efeitos físicos da TRIZ. Sob a orientação teórica da KBV, a tese reforça a compreensão de que a gestão eficiente do conhecimento é determinante para a inovação. Ao propor uma arquitetura híbrida que combina IA e modelagem ontológica, a pesquisa não apenas cria um instrumento técnico de extração de conhecimento, mas também amplia o acesso e o compartilhamento de saberes, consolidando uma ponte entre ciência, tecnologia e impacto social, em alinhamento com os ODS. A Figura 29 apresenta o infográfico que sumariza os impactos da pesquisa na sociedade.

Figura 29

Impactos da pesquisa na sociedade



Nota. Dados da pesquisa (2025).

7.2 Avaliação dos Produtos Técnico-Tecnológicos da Tese segundo critérios da CAPES

No âmbito da pesquisa, foram concebidos e validados dois PTTs: (i) uma ontologia semântica funcional baseada nos efeitos físicos da TRIZ, pioneira em língua portuguesa e estruturada para outros idiomas; e (ii) um método híbrido de mineração de inteligência técnica em patentes, integrando LLMs e Lógica de Decisão Derivativa, com foco na extração do relacionamento semântico ternário T–O–EF.

À luz dos critérios de avaliação da CAPES, os PTTs desta tese atendem integralmente ao requisito de Aderência e apresentam elevado impacto, alta aplicabilidade, forte caráter de Inovação e elevada complexidade. Os artefatos desenvolvidos materializam o conhecimento científico em soluções concretas, replicáveis e socialmente relevantes, consolidando a contribuição da pesquisa para a ciência, para a gestão da inovação e para o desenvolvimento tecnológico. A Tabela 34 apresenta a síntese avaliativa dos Produtos Técnico-Tecnológicos segundo os critérios da CAPES, evidenciando o atendimento aos parâmetros estabelecidos.

Tabela 34

Síntese Avaliativa dos Produtos Técnico-Tecnológicos segundo os Critérios da CAPES

Critério CAPES	Análise Sintética dos Produtos Técnico-Tecnológicos (PTTs)
Aderência	Os PTTs (ontologia TRIZ semântica funcional e método de mineração textual de patentes) derivam diretamente das linhas de pesquisa, objetivos e projetos estruturantes da tese. Estão intrinsecamente vinculados à articulação entre KBV, DSR, mineração textual, TRIZ e IA, configurando resultados centrais do percurso científico do programa de pós-graduação.
Impacto	Apresentam impacto multidimensional: acadêmico, ao avançar o estado da arte em mineração de patentes em português; prático/organizacional, ao apoiar gestão da inovação e prospecção tecnológica; social, ao democratizar o acesso ao conhecimento tecnológico em contextos lusófonos; político, ao subsidiar políticas públicas de CT&I; e educacional, ao viabilizar uso pedagógico em pós-graduação.
Aplicabilidade	Alta aplicabilidade, com uso prático, replicável e escalável. O método e a ontologia são disponibilizados em repositórios públicos, permitindo adoção

	por empresas, instituições públicas, centros de P&D e pesquisadores. Possibilitam a extração automatizada de relações Tarefa–Objeto–Efeito Físico (T–O–EF), convertendo grandes volumes de patentes em conhecimento acionável.
Inovação	Elevado grau de inovação metodológica e tecnológica. Propõe arquitetura híbrida original que integra LLMs, Lógica de Decisão Derivativa e ontologia semântica. A ontologia em língua portuguesa é inédita, especialmente em um contexto de idioma de baixo recurso computacional. O desempenho alcançado (73,26% de Precisão e EM) demonstra competitividade frente a abordagens de última geração.
Complexidade	Produto de alta complexidade, envolvendo múltiplos conhecimentos, atores e relações. Integra fundamentos da KBV, TRIZ e capacidade absortiva, métodos de DSR, técnicas avançadas de PLN, LLMs, lógica formal e validação por especialistas. Articula domínios como gestão da inovação, ciência da informação, linguística computacional e engenharia do conhecimento.

Nota. Elaborado pela Autora (2025).

7.3 Limitações da Pesquisa e Sugestões de Pesquisas Futuras

A pesquisa apresenta um conjunto de limitações que podem ser categorizadas em quatro áreas principais: (i) questões metodológicas das RSLs, (ii) desafios tecnológicos, linguísticos e ontológicos que afetam o desenvolvimento do método, e (iii) desafios relacionados à etapa de validação.

Nas questões metodológicas das RSLs, eventuais vieses podem ter ocorrido no escopo do estudo, bem como nas questões metodológicas necessárias para assegurar a qualidade da pesquisa. Nesse sentido, a definição da estratégia de busca pode ter utilizado termos excessivamente amplos ou restritivos, levando à omissão de estudos relevantes. Além disso, os critérios de seleção das publicações podem ter reduzido a abrangência dos resultados, ou as bases consultadas podem ter favorecido determinados estudos em detrimento de outros. Tais limitações são inerentes a qualquer RSL (Ang, 2018; Noor et al., 2023; Shaheen et al., 2023).

No desenvolvimento do método, os desafios tecnológicos, linguísticos e ontológicos são especialmente relevantes. Os documentos de patente extraídos da base do INPI fornecem apenas o título e o resumo, enquanto a descrição técnica e as reivindicações estão disponíveis apenas em formato PDF, o que limita a análise e dificulta a exploração de campos textuais mais ricos.

Muitos algoritmos ainda apresentam resultados variáveis entre diferentes domínios técnicos, o que evidencia a necessidade de desenvolver modelos de IA mais adaptáveis. Soma-se a essa limitação o fato de a análise ocorrer em textos em língua portuguesa, cuja carência de modelos robustos de PLN restringe o desempenho das técnicas e a replicabilidade dos resultados.

Quanto à ontologia, que serve como base semântica e lexical para o método de mineração, ainda são necessários expansão e refinamento, a fim de incorporar novos termos alinhados ao vocabulário contemporâneo e à língua portuguesa. Também são necessárias considerações quanto à integração de bases de conhecimento, visando à atualização contínua e à evolução semântica dos termos.

Na etapa de validação por especialistas, o processo de avaliação humana está sujeito ao julgamento individual, influenciado por fatores como nível de formação, experiência prévia e até desatenção nas respostas. Além disso, o anonimato do processo pode favorecer certo descuido.

A validação de cada patente foi realizada a partir da resposta de apenas um especialista, o que pode introduzir viés individual. O ideal, embora limitado por questões práticas, como o número reduzido de profissionais disponíveis e o tempo necessário, seria que cada documento fosse avaliado por, no mínimo, dois especialistas, a fim de verificar possíveis disparidades de interpretação.

A ausência de uma segunda rodada de avaliação, ou de uma etapa de iteração, é um fator que compromete ajustes e validações mais refinadas, uma vez que impede que os resultados iniciais retroalimentem o modelo. Esta restrição se torna ainda mais relevante considerando que a análise manual de patentes, a base para a validação, é um processo demorado e exige profissionais especializados, cuja disponibilidade é naturalmente restrita. Adicionalmente, o tamanho reduzido da amostra analisada constitui uma limitação para a generalização dos resultados. Nesse sentido, para garantir que as conclusões sejam robustas e não influenciadas pela escolha específica da amostra, sugere-se a aplicação futura de validação cruzada, na qual o processo é repetido com diferentes subamostras para validar o desempenho de forma mais abrangente (Miric et al., 2023).

Como direções para pesquisas futuras e perspectivas, podem ser identificados três eixos

principais: (i) o aprimoramento do método, voltado ao aumento da precisão, eficiência e interpretabilidade do processo de mineração; (ii) o enriquecimento da ontologia, com ênfase em sua atualização e expansão contínuas; e (iii) a ampliação do número de patentes analisadas e validadas.

No aprimoramento e refinamento metodológico, propõe-se o desenvolvimento do componente de derivação (Lógica de Decisão Derivativa) por meio do ajuste de hiperparâmetros (como temperatura e número de camadas do modelo), de forma a elevar a precisão na validação de novos termos das categorias T, O e EF, assegurando a preservação da coerência semântica com a ontologia original. Além disso, recomenda-se o ajuste dos modelos híbridos de IA, com o objetivo de aprimorar a precisão, a eficiência e a interpretabilidade na recuperação de informações tecnológicas, conforme recomendado por Jiang et al. (2025) e Trapp e Warschat (2025).

No eixo de expansão da ontologia, propõe-se sua integração a bases complementares, como tesauros técnicos, dicionários de sinônimos e antônimos, classificadores internacionais de patentes, artigos científicos, teses e dissertações. Essa integração visa promover o enriquecimento e o refinamento lexical contínuo. O enriquecimento semântico, por sua vez, pode ser alcançado por meio do desenvolvimento e da expansão de redes semânticas, explorando relações como sinônimos, hiperônimos e hipônimos, e pela incorporação de ontologias específicas de domínio, capazes de oferecer uma estrutura mais sensível ao contexto técnico.

Quanto ao *corpus* textual analisado, recomenda-se a exploração de fontes textuais mais ricas, incorporando campos mais extensos das patentes, como a descrição técnica e as reivindicações. Essa ampliação pode proporcionar maior contexto semântico, permitindo comparações de desempenho entre diferentes seções, como título, resumo e descrição, e contribuindo para uma análise mais robusta e contextualizada.

REFERÊNCIAS

Abbas, A., Zhang, L., & Khan, S. U. (2014). A literature review on the state-of-the-art in patent analysis. *World Patent Information*, *37*, 3–13. https://doi.org/10.1016/j.wpi.2013.12.006

Aggarwal, A., Sharma, C., Jain, M., Jain, A., Aggarwal, A., Sharma, C., Jain, M., & Jain, A. (2018). Semi Supervised Graph Based Keyword Extraction Using Lexical Chains and Centrality Measures. *Computación y Sistemas*, *22*(4), 1307–1315. https://doi.org/10.13053/cys-22-4-3077

Ali, A., Humayun, M. A., Silva, L. C. D., & Abas, P. E. (2025). Optimizing Patent Prior Art Search: An Approach Using Patent Abstract and Key Terms. *Information*, *16*(2), 145. https://doi.org/10.3390/info16020145

Ali, S., Li, G., Yang, P., Hussain, K., & Latif, Y. (2020). Unpacking the importance of intangible skills in new product development and sustainable business performance; strategies for marketing managers. *PLOS ONE*, *15*(9), e0238743. https://doi.org/10.1371/journal.pone.0238743

Almeida, R., Campos, R., Jorge, A., & Nunes, S. (2024). *Indexing Portuguese NLP Resources with PT-Pump-Up* (arXiv:2401.15400). arXiv. https://doi.org/10.48550/arXiv.2401.15400

Althammer, S., Buckley, M., Hofstätter, S., & Hanbury, A. (2021). *Linguistically Informed Masking for Representation Learning in the Patent Domain*. *2909*.

An, L. T. N., Matsuura, Y., & Oshima, N. (2024). Literature Review: Advanced Computational Tools for Patent Analysis. Em B. Alareeni & A. Hamdan (Orgs.), *Technology and Business Model Innovation: Challenges and Opportunities* (p. 483–494). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-55911-2_47

An, X., Li, J., Xu, S., Chen, L., & Sun, W. (2021). An improved patent similarity measurement based on entities and semantic relations. *Journal of Informetrics*, *15*(2), 101135. https://doi.org/10.1016/j.joi.2021.101135

Ang, L. (2018). Methodological reflections on the use of systematic reviews in early childhood research. *Journal of Early Childhood Research*, *16*(1), 18–31. https://doi.org/10.1177/1476718X17750206

Angeloni, M. T. (2003). Intervening elements in decision making. *Ciência da Informação*, *32*(1), 17–22. https://doi.org/10.1590/S0100-19652003000100002

Antons, D., Grünwald, E., Cichy, P., & Salge, T. O. (2020). The application of text mining methods in innovation research: Current state, evolution patterns, and development priorities. *R&D Management*, *50*(3), 329–351. https://doi.org/10.1111/radm.12408

Antons, D., Kleer, R., & Salge, T. O. (2016). Mapping the Topic Landscape of JPIM, 1984–2013: In Search of Hidden Structures and Development Trajectories. *Journal of Product Innovation Management*, *33*(6), 726–749. https://doi.org/10.1111/jpim.12300

Aras, H., Dessi, R., Saad, F., & Zhang, L. (2024). Bridging the Innovation Gap: Leveraging Patent Information for Scientists by Constructing a Patent-centric Knowledge Graph. *CEUR Workshop Proceedings*, *3697*, 61–67.

Aristodemou, L., & Tietze, F. (2018). The state-of-the-art on Intellectual Property Analytics (IPA): A literature review on artificial intelligence, machine learning and deep learning methods for analysing intellectual property (IP) data. *World Patent Information*, *55*, 37–51. https://doi.org/10.1016/j.wpi.2018.07.002

Aristodemou, L., Tietze, F., Athanassopoulou, N., & Minshall, T. (2017). *Exploring the future of patent analytics: A technology roadmapping approach*. University Cambridge.

https://www.ifm.eng.cam.ac.uk/insights/innovation-and-ip-management/exploring-the-future-of-patent-analytics/

Arts, S., Hou, J., & Gomez, J. C. (2021). Natural language processing to identify the creation and impact of new technologies in patent text: Code, data, and new measures. *Research Policy*, *50*(2), 104144. https://doi.org/10.1016/j.respol.2020.104144

Audretsch, D. B., & Feldman, M. P. (1996). R&D Spillovers and the Geography of Innovation and Production. *American Economic Review*, *86*(3), 630–640.

Aulive. (2025). *Production Inspiration*. Aulive. https://www.productioninspiration.com/

Ayaou, I., Chibane, H., Koch, S., & Cavallucci, D. (2025). *Leveraging Information Retrieval Pipelines for Inventive Design: Application in Efficient Lattice Structures Manufacturing*. *736 IFIP*, 321–329. Scopus. https://doi.org/10.1007/978-3-031-75923-9_21

Azevedo, M. M. (2005). *Portuguese: A linguistic introduction*. Cambridge University Press.

Baonza, M. del C. S. de F. (2010). *NeOn Methodology for Building Ontology Networks: Specification, scheduling and reuse* [Universidad Politécnica de Madrid]. https://oa.upm.es/3879/2/maria_del-_carmen_suarez_de_figueiroa_baonza.pdf

Becattini, N., Borgianni, Y., Cascini, G., & Rotini, F. (2015). *ARIZ85 and patent-driven knowledge support*. *131*, 291–302. https://doi.org/10.1016/j.proeng.2015.12.391

Behkami, N. A., & Daim, T. U. (2012). Research Forecasting for Health Information Technology (HIT), using technology intelligence. *Technological Forecasting and Social Change*, *79*(3), 498–508. https://doi.org/10.1016/j.techfore.2011.08.015

Belenzon, S. (2012). Cumulative Innovation and Market Value: Evidence from Patent Citations. *The Economic Journal*, *122*(559), 265–285. https://doi.org/10.1111/j.1468-0297.2011.02470.x

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. https://doi.org/10.1145/3442188.3445922

Berdyugina, D., & Cavallucci, D. (2020a). *Improvement of Automatic Extraction of Inventive Information with Patent Claims Structure Recognition*. *1229 AISC*, 625–637. https://doi.org/10.1007/978-3-030-52246-9_46

Berdyugina, D., & Cavallucci, D. (2020b). *Setting Up Context-Sensitive Real-Time Contradiction Matrix of a Given Field Using Unstructured Texts of Patent Contents and Natural Language Processing*. *597 IFIP*, 30–39. https://doi.org/10.1007/978-3-030-61295-5_3

Berdyugina, D., & Cavallucci, D. (2021). *Automatic Extraction of Potentially Contradictory Parameters from Specific Field Patent Texts*. *635 IFIP*, 150–161. https://doi.org/10.1007/978-3-030-86614-3_12

Berdyugina, D., & Cavallucci, D. (2022a). *Exploitation of Causal Relation for Automatic Extraction of Contradiction from a Domain-Restricted Patent Corpus*. *655 IFIP*, 86–95. https://doi.org/10.1007/978-3-031-17288-5_8

Berdyugina, D., & Cavallucci, D. (2022b). Natural Language Processing in assistance to Inventive Design activities. *Procedia CIRP*, *109*, 7–12. https://doi.org/10.1016/j.procir.2022.05.206

Berdyugina, D., & Cavallucci, D. (2023). Automatic extraction of inventive information out of patent texts in support of manufacturing design studies using Natural Languages Processing. *Journal of Intelligent Manufacturing*, *34*(5), 2495–2509. https://doi.org/10.1007/s10845-022-01943-y

Bianchi, M., Croce, A., Dell'Era, C., Di Benedetto, C. A., & Frattini, F. (2016). Organizing for Inbound Open Innovation: How External Consultants and a Dedicated R&D Unit Influence Product Innovation Performance. *Journal of Product Innovation Management*, *33*(4), 492–510. https://doi.org/10.1111/jpim.12302

Blume, M., Heidari, G., & Hewel, C. (2024). *Comparing Complex Concepts with Transformers: Matching Patent Claims Against Natural Language Text*. *3775*. https://doi.org/10.48550/arXiv.2407.10351

Brad, S. (2023). Mapping the Evolutionary Journey of TRIZ and Pioneering Its Next S-Curve in the Age of AI-Aided Invention. Em D. Cavallucci, P. Livotov, & S. Brad (Orgs.), *Towards AI-Aided Invention and Innovation* (p. 3–22). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-42532-5_1

Bregonje, M. (2005). Patents: A unique source for scientific technical information in chemistry related industry? *World Patent Information*, *27*(4), 309–315. https://doi.org/10.1016/j.wpi.2005.05.003

Cabrilo, S., & Dahms, S. (2018). How strategic knowledge management drives intellectual capital to superior innovation and market performance. *Journal of Knowledge Management*, *22*(3), 621–648. https://doi.org/10.1108/JKM-07-2017-0309

Cao, G., Luo, P., Wang, L., & Yang, X. (2016). *Key Technologies for Sustainable Design Based on Patent Knowledge Mining*. *39*, 97–102. Scopus. https://doi.org/10.1016/j.procir.2016.01.172

Cao, Y. S., Tan, J., & Wang, K. W. (2018). Analysis of technology research and development situation of China's grain and oil industry based on patent information. *Science and Technology Management Research*, *38*(8), 131–138. https://doi.or/10.3969/j.issn.1000-7695.2018.08.019

Carvalho, M. A., & Back, N. (2001). Uso dos conceitos fundamentais da TRIZ e do método dos princípios inventivos no desenvolvimento de produtos. *Anais do III Congresso Brasileiro de Gestão de Desenvolvimento de Produto*. III Congresso Brasileiro de Gestão de Desenvolvimento de Produto, Florianópolis, Brasil. https://www.researchgate.net/publication/255664510_uso_dos_conceitos_fundamentais_da_triz_e_do_metodo_dos_principios_inventivos_no_desenvolvimento_de_produtos

Cascini, G., & Russo, D. (2007). Computer-aided analysis of patents and search for TRIZ contradictions. *International Journal of Product Development*, *4*(1–2), 52–67. https://doi.org/10.1504/IJPD.2007.011533

Cascini, G., & Zini, M. (2011). *Computer-aided comparison of thesauri extracted from complementary patent classes as a means to identify relevant field parameters*. 555–566. https://doi.org/10.1007/978-3-642-15973-2_56

Cavallucci, D., Rousselot, F., & Zanni, C. (2011a). An ontology for TRIZ. *Procedia Engineering*, *9*, 251–160.

Cavallucci, D., Rousselot, F., & Zanni, C. (2011b). Using patents to populate an inventive design ontology. *Procedia Engineering*, *9*, 52–62. https://doi.org/10.1016/j.proeng.2011.03.100

Chan, C. K., Ng, K. W., Ang, M. C., Ng, C. Y., & Kor, A.-L. (2021). *Sustainable Product Innovation Using Patent Mining and TRIZ*. *13051*, 287–298. https://doi.org/10.1007/978-3-030-90235-3_25

Chan, E.-M., Kor, A.-L., Ng, K. W., Ang, M. C., & Wahab, A. N. A. (2021). A Conceptual Design Framework based on TRIZ Scientific Effects and Patent Mining. *International Journal of Advanced Computer Science and Applications*, *12*(12), 43–50. https://doi.org/10.14569/IJACSA.2021.0121206

Chandra, S., & Livotov, P. (2019). *Classification of TRIZ Inventive Principles and Sub-principles for Process Engineering Problems*. *572*, 314–327. https://doi.org/10.1007/978-3-030-32497-1_26

Chao, M.-H., Trappey, A., Wu, C.-T., & Su, Y.-A. (2021). Technology Mining for Intelligent Chatbot Development. Em *Transdisciplinary Engineering for Resilience: Responding to System Disruptions*. https://doi.org/10.3233/ATDE210090

Charan, J. (2014). Impact factor: Is this a true measure of quality? *International Journal of Medical Science and Public Health*, *3*(3), 246. https://doi.org/10.5455/ijmsph.2014.190420141

Chen, D. (2024). Challenges of Natural Language Processing from a Linguistic Perspective. *International Journal of Education and Humanities*, *13*(2), Artigo 2. https://doi.org/10.54097/hyapye19

Chen, J., Chen, J., Zhao, S., Zhang, Y., & Tang, J. (2020). Exploiting word embedding for heterogeneous topic model towards patent recommendation. *Scientometrics*, *125*(3), 2091–2108. https://doi.org/10.1007/s11192-020-03666-4

Chen, L., Xu, S., Shang, W., Zheng, W., Wei, C., & Xu, H.-Y. (2020). *What is Special about Patent Information Extraction?* 1st workshop on Extraction and Evaluation of Knowledge Entities from Scientific Documents at the ACM/IEEE JCDL2020, virtual conference. https://www.researchgate.net/publication/344036678_What_is_Special_about_Patent_Information_Extraction/link/5f4ecf4f92851c250b88b92f/download?_tp=eyJjb250ZXh0Ijp7ImZpcnN0UGFnZSI6InB1YmxpY2F0aW9uIiwicGFnZSI6InB1YmxpY2F0aW9uIn19

Chen, L., Xu, S., Zhu, L., Zhang, J., Yang, G., & Xu, H. (2022). A deep learning based method benefiting from characteristics of patents for semantic relation classification. *Journal of Informetrics*, *16*(3), 101312. https://doi.org/10.1016/j.joi.2022.101312

Chiarello, F., Cimino, A., Fantoni, G., & Dell'Orletta, F. (2018). Automatic users extraction from patents. *World Patent Information*, *54*, 28–38. https://doi.org/10.1016/j.wpi.2018.07.006

Choi, S., Kang, D., Lim, J., & Kim, K. (2012). A fact-oriented ontological approach to SAO-based function modeling of patents for implementing Function-based Technology Database. *Expert Systems with Applications*, *39*(10), 9129–9140. https://doi.org/10.1016/j.eswa.2012.02.041

Choi, S., Kim, H., Yoon, J., Kim, K., & Lee, J. Y. (2013). An SAO-based text-mining approach for technology roadmapping using patent information. *R and D Management*, *43*(1), 52–74. https://doi.org/10.1111/j.1467-9310.2012.00702.x

Choi, S., Park, H., Kang, D., Lee, J. Y., & Kim, K. (2012). An SAO-based text mining approach to building a technology tree for technology planning. *Expert Systems with Applications*, *39*(13), 11443–11455. https://doi.org/10.1016/j.eswa.2012.04.014

Choi, Y., Park, S., & Lee, S. (2021). Identifying emerging technologies to envision a future innovation ecosystem: A machine learning approach to patent data. *Scientometrics*, *126*(7), 5431–5476. https://doi.org/10.1007/s11192-021-04001-1

Chuprat, S., Novianto, E. H. D., Matsuura, Y., Mahdzir, A. M., & Harun, A. N. (2024). A closer look on patent analytics through systematic literature review. *Management Review Quarterly*. https://doi.org/10.1007/s11301-024-00452-x

Cohen, W., & Levinthal, D. (1990). Absorptive Capacity: A New Perspective on Learning and Innovation. *Administrative Science Quarterly*, *35*, 128–152. https://doi.org/10.2307/2393553

Cong, H., & Tong, L. H. (2008). Grouping of TRIZ Inventive Principles to facilitate automatic patent classification. *Expert Systems with Applications*, *34*(1), 788–795. https://doi.org/10.1016/j.eswa.2006.10.015

Costa, P. R. da, Pigola, A., Ramos, H. R., & Pedron, C. D. (2024). Estruturas de tese de doutorado para convergência científica, técnica, tecnológica e social. *Revista Ibero-Americana de Estratégia*, *23*(1), e26202–e26202. https://doi.org/10.5585/2024.26202

Costa, P. R. da, Ramos, H. R., & Pedron, C. D. (2019). Proposição de Estrutura Alternativa para Tese de Doutorado a Partir de Estudos Múltiplos. *Revista Ibero-Americana de Estratégia*, *18*(2), 155–170. https://doi.org/10.5585/riae.v18i2.2783

Cui, X., & Qian, L. (2022). Research on Patent Information Extraction Based on Deep Learning. Em H. Yang, R. Qiu, & W. Chen (Orgs.), *AI and Analytics for Public Health* (p. 291–302). Springer International Publishing. https://doi.org/10.1007/978-3-030-75166-1_21

Cunha, K. C. T., Volpato, G., & Pedron, C. D. (2023). Documentos de patente como fonte de informação: A perspectiva dos pesquisadores. *Anais do XV Encontro Acadêmico de Propriedade Intelectual, Inovação e Desenvolvimento: "Propriedade Intelectual como Instrumento de estímulo ao Desenvolvimento Sustentável*, 124–129. https://www.gov.br/inpi/pt-br/servicos/a-academia/eventos-academicos/enapid/enapid-2023/arquivos/enapid-2023_final.pdf

de Weck, O. L. (2022). Patents and Intellectual Property. Em O. L. De Weck (Org.), *Technology Roadmapping and Development: A Quantitative Approach to the Management of Technology* (p. 119–152). Springer International Publishing. https://doi.org/10.1007/978-3-030-88346-1_5

Deng, N., Chen, X., Ruan, O., Wang, C., Ye, Z., & Tian, J. (2018). PaEffExtr: A Method to Extract Effect Statements Automatically from Patents. Em L. Barolli & O. Terzo (Orgs.), *Complex, Intelligent, and Software Intensive Systems* (p. 667–676). Springer International Publishing. https://doi.org/10.1007/978-3-319-61566-0_62

Deng, N., Wang, C., Zhang, M., Ye, Z., Xiao, L., Tian, J., Li, D., & Chen, X. (2018). A Semi-Automatic Annotation Method of Effect Clue Words for Chinese Patents Based on Co-Training. *Revista Internacional de Data Warehousing e Mineração*, *14*(4), 1–19. https://doi.org/10.4018/IJDWM.2018100101

Dessi, R., Aras, H., & Zhang, L. (2023). DeepKEA: Employing Deep Learning Models for Keyword Extraction from Patent Documents. *CEUR Workshop Proceedings*, 41–46. https://ceur-ws.org/Vol-3594/paper3.pdf

Dewulf, S. (2011). Directed variation of properties for new or improved function product DNA – A base for connect and develop. *Procedia Engineering*, *9*, 646–652. https://doi.org/10.1016/j.proeng.2011.03.150

Dewulf, S., & Childs, P. R. N. (2023). *Innovation Logic: Benefits of a TRIZ-Like Mind in AI Using Text Analysis of Patent Literature*. *682*, 95–102. Scopus. https://doi.org/10.1007/978-3-031-42532-5_7

Ding, Z., Jiang, S., Ng, F., & Zhu, M. (2017). A new TRIZ-based patent knowledge management system for construction technology innovation. *Journal of Engineering, Design and Technology*, *15*(4), 456–470. https://doi.org/10.1108/JEDT-03-2016-0017

Ding, Z., & Ma, J. (2014). *An exploration study of construction innovation principles: Comparative analysis of construction scaffold and template patents*. 843–850. https://doi.org/10.1007/978-3-642-35548-6_86

Dintzner, J.-P., & Van Thieleny, J. (1991). Image handling at the European Patent Office: BACON and first page. *World Patent Information*, *13*(3), 152–154. https://doi.org/10.1016/0172-2190(91)90070-L

Donald, K. E., Kabir, K. M. M., & Donald, W. A. (2018). Tips for reading patents: A concise introduction for scientists. *Expert Opinion on Therapeutic Patents*, *28*(4), 277–280. https://doi.org/10.1080/13543776.2018.1438409

Dresch, A., Lacerda, D. P., & Miguel, P. A. C. (2015). A Distinctive Analysis of Case Study, Action Research and Design Science Research. *Review of Business Management*, 1116–1133. https://doi.org/10.7819/rbgn.v17i56.2069

Du, R., Sun, J., Miao, R., & Zhang, D. (2025). AI-Aided Resource Mining Method for Idealization-Driven Product Innovation. *IFIP Advances in Information and Communication Technology*, *735 IFIP*, 147–164. https://doi.org/10.1007/978-3-031-75919-2_9

Dybå, T., Dingsøyr, T., & Hanssen, G. (2007). Applying Systematic Reviews to Diverse Study Types: An Experience Report. *Proceedings - 1st International Symposium on Empirical Software Engineering and Measurement, ESEM 2007*, 234. https://doi.org/10.1109/ESEM.2007.59

Ekmekci, I., & Nebati, E. E. (2019). Triz Methodology and Applications. *Procedia Computer Science*, *158*, 303–315. https://doi.org/10.1016/j.procs.2019.09.056

European Patent Office. (2025a). *Bibliographic coverage in Espacenet and OPS*. https://www.epo.org/en/searching-for-patents/data/coverage

European Patent Office. (2025b). *Cooperative Patent Classification (CPC)*. Cooperative Patent Classification (CPC). https://www.epo.org/en/searching-for-patents/helpful-resources/first-time-here/classification/cpc?utm_source=chatgpt.com

Fall, C. J., Törcsvári, A., Benzineb, K., & Karetka, G. (2003). Automated categorization in the international patent classification. *Fórum ACM SIGIR*, *37*(1), 10–25. https://doi.org/10.1145/945546.945547

Feldman, R., Ronen, Sanger, & James. (2007). *The text mining handbook: Advanced approaches in analyzing unstructured data* (Vol. 34). Cambridge University Press.

Fink, T., Andersson, L., & Hanbury, A. (2021). Detecting Multi Word Terms in patents the same way as entities. *World Patent Information*, *67*, 102078. https://doi.org/10.1016/j.wpi.2021.102078

Florescu, C., & Caragea, C. (2017). PositionRank: An Unsupervised Approach to Keyphrase Extraction from Scholarly Documents. *In: Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL 2017)*. https://doi.org/10.18653/v1/p17-1102

Fomenkova, M., Korobkin, D., & Fomenkov, S. (2017). *Extraction of physical effects based on the semantic analysis of the patent texts*. *754*, 73–87. https://doi.org/10.1007/978-3-319-65551-2_6

Gadd, K. (2011). *TRIZ for Engineers: Enabling Inventive Problem Solving*. John Wiley & Sons.

Galvão, M. C. B., & Ricarte, I. L. M. (2019). Systematic Literature review: Concept, production and publication. *Logeion: Filosofia da Informação*, *6*(1), 57–73. https://doi.org/10.21728/logeion.2019v6n1.p57-73

Gao, Y., & Zhu, Y. (2015). Research on Dynamic Capabilities and Innovation Performance in the Chinese Context: A Theory Model-Knowledge Based View. *Open Journal of Business and Management*, *03*(04), Artigo 04. https://doi.org/10.4236/ojbm.2015.34035

Geng, B. (2021). Open Relation Extraction in Patent Claims with a Hybrid Network. *Wireless Communications and Mobile Computing*, *2021*, 1–7. https://doi.org/10.1155/2021/5547281

Gerken, J. M., & Moehrle, M. G. (2012). A new instrument for technology monitoring: Novelty in patents measured by semantic patent analysis. *Scientometrics*, *91*(3), 645–670. https://doi.org/10.1007/s11192-012-0635-7

Ghoula, N., Khelif, K., & Dieng-Kuntz, R. (2007). Supporting Patent Mining by using Ontology-based Semantic Annotations. *IEEE/WIC/ACM International Conference on Web Intelligence (WI'07)*, 435–438. https://doi.org/10.1109/WI.2007.45

Giordano, V., Consoloni, M., Chiarello, F., & Fantoni, G. (2024). Towards the extraction of semantic relations in design with natural language processing. *Proceedings of the Design Society*, *4*, 2059–2068. https://doi.org/10.1017/pds.2024.208

Giordano, V., Puccetti, G., Chiarello, F., Pavanello, T., & Fantoni, G. (2023). Unveiling the inventive process from patents by extracting problems, solutions and advantages with natural language processing. *Expert Systems with Applications*, *229*, 120499. https://doi.org/10.1016/j.eswa.2023.120499

GO FAIR International Support & Coordination Office. (2025). *FAIR Principles*. GO FAIR. https://www.go-fair.org/fair-principles/

Gongchang, R., Qi, L., & Fenghai, Y. (2014). *On classification and extraction of deep knowledge in patents based on TRIZ Theory*. 666–670. https://doi.org/10.1109/ISDEA.2014.154

Granovetter, M. S. (1973). The Strength of Weak Ties. *American Journal of Sociology*, *78*(6), 1360–1380. https://doi.org/10.1086/225469

Grant, R. M. (1996). Toward a knowledge-based theory of the firm. *Strategic Management Journal*, *17*(S2), 109–122. https://doi.org/10.1002/smj.4250171110

Grant, R. M. (1997). The knowledge-based view of the firm: Implications for management practice. *Long Range Planning*, *30*(3), 450–454. https://doi.org/10.1016/S0024-6301(97)00025-3

Grant, R., & Phene, A. (2022). The knowledge based view and global strategy: Past impact and future potential. *Global Strategy Journal*, *12*(1), 3–30. https://doi.org/10.1002/gsj.1399

Guarino, G., Samet, A., & Cavallucci, D. (2020). *Summarization as a Denoising Extraction Tool*. *597 IFIP*, 77–87. https://doi.org/10.1007/978-3-030-61295-5_7

Guarino, G., Samet, A., & Cavallucci, D. (2021). *Patent Specialization for Deep Learning Information Retrieval Algorithms*. *635 IFIP*, 162–169. https://doi.org/10.1007/978-3-030-86614-3_13

Guarino, G., Samet, A., & Cavallucci, D. (2022). PaTRIZ: A framework for mining TRIZ contradictions in patents. *Expert Systems with Applications*, *207*. https://doi.org/10.1016/j.eswa.2022.117942

Guarino, G., Samet, A., & Cavallucci, D. (2024). *SynCRF: Syntax-Based Conditional Random Field for TRIZ Parameter Minings*. *3*, 890–897. https://doi.org/10.5220/0012411300003636

Guarino, G., Samet, A., Nafi, A., & Cavallucci, D. (2020). *SummaTRIZ : Summarization Networks for Mining Patent Contradiction*. 979–986. https://doi.org/10.1109/ICMLA51294.2020.00159

Han, S. H., Kim, H.-J., Cho, K., Kim, M. K., Kim, H., & Park, S.-H. (2006). Research planning methodology for technology fusion in construction. *2006 Proceedings of the 23rd International Symposium on Robotics and Automation in Construction, ISARC 2006*, 15–18. https://doi.org/10.22260/ISARC2006/0005

Hansen, M. T. (1999). The Search-Transfer Problem: The Role of Weak Ties in Sharing Knowledge across Organization Subunits. *Administrative Science Quarterly*, *44*(1), 82–111. https://doi.org/10.2307/2667032

He, C., Tan, R., Peng, Q., Shi, F., Shao, P., & Li, X. (2022). Improvement of Technological Innovation of SMEs Using Patent Knowledge. *Computer-Aided Design and Applications*, *19*, 936–951. https://doi.org/10.14733/cadaps.2022.936-951

Heisig, P., Ogaza, M. A., & Hamraz, B. (2020). Information and knowledge assessment – Results from a multinational automotive company. *International Journal of Information Management*, *54*, 102137. https://doi.org/10.1016/j.ijinfomgt.2020.102137

Helmers, L., Horn, F., Biegler, F., Oppermann, T., & Müller, K.-R. (2019). Automating the search for a patent's prior art with a full text similarity search. *PLOS ONE*, *14*(3), e0212103. https://doi.org/10.1371/journal.pone.0212103

Hevner, A., R, A., March, S., T, S., Park, Park, J., Ram, & Sudha. (2004). Design Science in Information Systems Research. *Management Information Systems Quarterly*, *28*, 75.

Hevner, A., vom Brocke, J., & Maedche, A. (2019). Roles of Digital Innovation in Design Science Research. *Business & Information Systems Engineering*, *61*(1), 3–8. https://doi.org/10.1007/s12599-018-0571-z

Higgins, J. P. T., Tomás, J., Chandler, J., Cumpston, M., Li, T., Page, M. J., & Welch, V. A. (2023). *Cochrane Handbook for Systematic Reviews of Interventions* (6.4). Cochrane. https://training.cochrane.org/handbook/current

Hmina, K., Sallaou, M., Ait Taleb, A., & Lasri, L. (2019). TRIZ The theory of Inventif Problem Solving State of the art. *2019 5th International Conference on Optimization and Applications (ICOA)*, 1–7. https://doi.org/10.1109/ICOA.2019.8727620

Hou, W., Liu, T., Li, B., Hong, Z., & Haiyang, W. (2024). Patent-Based Technology Efficacy Information Extraction in Product Innovation Design. Em J. Tan, Y. Liu, H.-Z. Huang, J. Yu, & Z. Wang (Orgs.), *Advances in Mechanical Design* (p. 419–427). Springer Nature. https://doi.org/10.1007/978-981-97-0922-9_26

Hu, J., Li, S., Yao, Y., Yu, L., Yang, G., & Hu, J. (2018). Patent Keyword Extraction Algorithm Based on Distributed Representation for Patent Classification. *Entropy*, *20*(2). https://doi.org/10.3390/e20020104

Huang, Z., Guo, X., Liu, Y., Zhao, W., & Zhang, K. (2023). A smart conflict resolution model using multi-layer knowledge graph for conceptual design. *Advanced Engineering Informatics*, *55*. Scopus. https://doi.org/10.1016/j.aei.2023.101887

Huang, Z., & Xie, Z. (2022). A patent keywords extraction method using TextRank model with prior public knowledge. *Complex & Intelligent Systems*, *8*(1), 1–12. https://doi.org/10.1007/s40747-021-00343-8

Hwang, S.-Y., Shin, D.-J., & Kim, J.-J. (2022). Systematic Review on Identification and Prediction of Deep Learning-Based Cyber Security Technology and Convergence Fields. *Symmetry*, *14*(4), Artigo 4. https://doi.org/10.3390/sym14040683

Ilevbare, I. M., Probert, D., & Phaal, R. (2013). A review of TRIZ, and its benefits and challenges in practice. *Technovation*, *33*(2), 30–37. https://doi.org/10.1016/j.technovation.2012.11.003

INPI. (2024, abril 29). *Instituto Nacional da Propriedade Industrial*. Instituto Nacional da Propriedade Industrial. https://www.gov.br/inpi/pt-br/copy2_of_nova-home-page

Jang, H., Jeong, Y., & Yoon, B. (2021). TechWord: Development of a technology lexical database for structuring textual technology information based on natural language processing. *Expert Systems with Applications*, *164*, 114042. https://doi.org/10.1016/j.eswa.2020.114042

Jang, H., Park, S., & Yoon, B. (2022). Exploring Technology Opportunities Based on User Needs: Application of Opinion Mining and SAO Analysis. *Engineering Management Journal*, 1–14. https://doi.org/10.1080/10429247.2022.2050130

Jang, H., & Yoon, B. (2021). TechWordNet: Development of semantic relation for technology information analysis using F-term and natural language processing. *Information Processing & Management*, *58*(6), 102752. https://doi.org/10.1016/j.ipm.2021.102752

Jarrar, Y. F. (2002). Knowledge management: Learning for organisational experience. *Managerial Auditing Journal*, *17*(6), 322–328. https://doi.org/10.1108/02686900210434104

Jeon, S. H., Lee, H. J., Park, J., & Cho, S. (2024). Building knowledge graphs from technical documents using named entity recognition and edge weight updating neural network with triplet loss for entity normalization. *Intelligent Data Analysis*, *28*(1), 331–355. https://doi.org/10.3233/IDA-227129

Jiang, Atherton, & Sorce. (2023). Extraction and linking of motivation, specification and structure of inventions for early design use. *Journal of Engineering Design*, *34*(5–6), 411–436. https://doi.org/10.1080/09544828.2023.2227934

Jiang, J., Ying, F., & Dhuny, R. (2025). Unveiling Technological Evolution with a Patent-Based Dynamic Topic Modeling Framework: A Case Study of Advanced 6G Technologies. *Applied Sciences*, *15*(7), 3783. https://doi.org/10.3390/app15073783

Jiang, M., & Shang, J. (2020). *Scientific Text Mining and Knowledge Graphs*. 3537–3538. https://doi.org/10.1145/3394486.3406465

Jing, L., Yang, J., Ma, J., Xie, J., Li, J., & Jiang, S. (2023). An integrated implicit user preference mining approach for uncertain conceptual design decision-making: A pipeline inspection trolley design case study. *Knowledge-Based Systems*, *270*, 110524. https://doi.org/10.1016/j.knosys.2023.110524

Joseph, S., Sedimo, K., Kaniwa, F., Hlomani, H., & Letsholo, K. (2016). Natural Language Processing: A Review. *Natural Language Processing: A Review*, *6*, 207–210.

Joshi, U., Hedaoo, M., Fatnani, P., Bansal, M., & More, V. (2022). Patent Classification with Intelligent Keyword Extraction. *2022 6th International Conference On Computing, Communication, Control And Automation (ICCUBEA*, 1–7. https://doi.org/10.1109/ICCUBEA54992.2022.10010888

Kaliteevskii, V., Deder, A., Peric, N., & Chechurin, L. (2020). *Conceptual Semantic Analysis of Patents and Scientific Publications Based on TRIZ Tools*. *597 IFIP*, 54–63. https://doi.org/10.1007/978-3-030-61295-5_5

Kaliteevskii, V., Deder, A., Peric, N., & Chechurin, L. (2021). *Concept Extraction Based on Semantic Models Using Big Amount of Patents and Scientific Publications Data*. *635 IFIP*, 141–149. https://doi.org/10.1007/978-3-030-86614-3_11

Kang, J., Souili, A., & Cavallucci, D. (2018). *Text Simplification of Patent Documents: 18th International TRIZ Future Conference, TFC 2018, Strasbourg, France, October 29–31, 2018, Proceedings* (p. 225–237). https://doi.org/10.1007/978-3-030-02456-7_19

Kashyap, G. (2021). Multilingual NLP: Techniques for Creating Models that Understand and Generate Multiple Languages with Minimal Resources. *International Journal of Scientific Research in Engineering and Management (IJSREM)*, *5*(5). https://doi.org/10.55041/IJSREM7648

Katz, R., & Allen, T. J. (1982). Investigating the Not Invented Here (NIH) syndrome: A look at the performance, tenure, and communication patterns of 50 R & D Project Groups. *R&D Management*, *12*(1), 7–20. https://doi.org/10.1111/j.1467-9310.1982.tb00478.x

Khadilkar, K., Kulkarni, S., & Venkatraman, S. (2019). A Knowledge Graph Based Approach for Automatic Speech and Essay Summarization. *2019 IEEE 5th International Conference for Convergence in Technology (I2CT)*, 1–6. https://doi.org/10.1109/I2CT45611.2019.9033908

Khode, A. (2019). Effect of Technical Domains and Patent Structure on Patent Information Retrieval. *International Journal of Engineering and Advanced Technology*, *9*(1). https://doi.org/10.35940/ijeat.A1922.109119

Ki, W., & Kim, K. (2017). Generating Information Relation Matrix Using Semantic Patent Mining for Technology Planning: A Case of Nano-Sensor. *IEEE Access*, *5*, 26783–26797. https://doi.org/10.1109/ACCESS.2017.2771371

Kim, Choi, Park, & Jang. (2018). Patent Keyword Extraction for Sustainable Technology Management. *Sustainability*, *10*(4), 1287. https://doi.org/10.3390/su10041287

Kim, D., Kim, N., & Kim, W. (2018). The effect of patent protection on firms' market value: The case of the renewable energy sector. *Renewable and Sustainable Energy Reviews*, *82*, 4309–4319. https://doi.org/10.1016/j.rser.2017.08.001

Kim, H., & Kim, K. (2012). Causality-based function network for identifying technological analogy. *Expert Systems with Applications*, *39*(12), 10607–10619. https://doi.org/10.1016/j.eswa.2012.02.156

Kim, J., Choi, J., Park, S., & Jang, D. (2018). Patent Keyword Extraction for Sustainable Technology Management. *Sustainability*, *10*, 1287. https://doi.org/10.3390/su10041287

Kim, Joung, & Kim. (2018). Semi-automatic extraction of technological causality from patents. *Computers and Industrial Engineering*, *115*, 532–542. https://doi.org/10.1016/j.cie.2017.12.004

Kim, K., Park, K., & Lee, S. (2019). Investigating technology opportunities: The use of SAOx analysis. *Scientometrics*, *118*(1), 45–70.

Kim, Kim, Lee, Lim, & Moon. (2009). Application of TRIZ creativity intensification approach to chemical process safety. *Journal of Loss Prevention in the Process Industries*, *22*(6), 1039–1043. https://doi.org/10.1016/j.jlp.2009.06.015

Kim, Park, & Lee. (2019). Investigating technology opportunities: The use of SAOx analysis. *Scientometrics*, *118*(1), 45–70. https://doi.org/10.1007/s11192-018-2962-9

Kim, S., Park, I., & Yoon, B. (2020). SAO2Vec: Development of an algorithm for embedding the subject–action–object (SAO) structure using Doc2Vec. *PLOS ONE*, *15*(2), e0227930. https://doi.org/10.1371/journal.pone.0227930

Kim, Y. G., Suh, J. H., & Park, S. C. (2008). Visualization of patent analysis for emerging technology. *Expert Systems with Applications*, *34*(3), 1804–1812. https://doi.org/10.1016/j.eswa.2007.01.033

Kim, & Yoon. (2022). Multi-document summarization for patent documents based on generative adversarial network. *Expert Systems with Applications*, *207*, 117983. https://doi.org/10.1016/j.eswa.2022.117983

Kitamura, Y., Taniguchi, K., & Kato, S. (2024). *Patent analysis using an ontology of qualities of inorganic materials based on context-dependency*. Joint Ontology Workshops. https://www.semanticscholar.org/paper/Patent-analysis-using-an-ontology-of-qualities-of-Kitamura-Taniguchi/d67bbc44570c1e12177f734d1095c0eab4af6c2d

Kogut, B., & Zander, U. (1996). What Firms Do? Coordination, Identity, and Learning. *Organization Science*, *7*(5), 502–518. https://doi.org/10.1287/orsc.7.5.502

Korobkin, & Fomenkov. (2018). Method of detection of technical functions performed by physical effects. *IOP Conf. Series: Earth and Environmental Science*. https://doi.org/10.1088/1755-1315/194/2/022014

Korobkin, Fomenkov, & Golovanchikov. (2018). Method of identification of patent trends based on descriptions of technical functions. *Journal of Physics: Conference Series*, *1015*(3), 032065. https://doi.org/10.1088/1742-6596/1015/3/032065

Korobkin, Fomenkov, & Kolesnikov. (2018). The method for detecting the dependencies between technical functions and physical effects. *Proceedings of the International Conferences on Big Data Analytics, Data Mining and Computational Intelligence 2018*, 225–228. https://www.researchgate.net/publication/329881212_the_method_for_detecting_the_dependencies_between_technical_functions_and_physical_effects

Korobkin, Fomenkov, & Kravets. (2017). *Extraction of physical effects practical applications from patent database*. *2018-January*, 1–5. https://doi.org/10.1109/IISA.2017.8316402

Korobkin, Fomenkov, & Kravets. (2018). Methods for Extracting the Descriptions of Sci-Tech Effects and Morphological Features of Technical Systems from Patents. *2018 9th International Conference on Information, Intelligence, Systems and Applications (IISA)*, 1–4. https://doi.org/10.1109/IISA.2018.8633624

Korobkin, Fomenkov, Kravets, Kolesnikov, & Dykov. (2015). *Three-steps methodology for patents prior-art retrieval and structured physical knowledge extracting*. *535*, 124–136. https://doi.org/10.1007/978-3-319-23766-4_10

Korobkin, Shabanov, Fomenkov, & Golovanchikov. (2019). *Construction of a Matrix "Physical Effects – Technical Functions" on the Base of Patent Corpus Analysis*. *1084*, 52–68. https://doi.org/10.1007/978-3-030-29750-3_5

Krasnov, F., Shvartsman, M., & Dimentov, A. (2022). Comparing text corpora via topic modelling. *International Journal of Data Mining, Modelling and Management*, *14*(3), 203–216. https://doi.org/10.1504/IJDMMM.2022.125259

Kraus, B., Matzke, S., Welzbacher, P., & Kirchner, E. (2022). Utilizing a graph data structure to model physical effects and dependencies between different physical variables for the

systematic identification of sensory effects in design elements. *DS 119: Proceedings of the 33rd Symposium Design for X (DFX2022)*, 1–10. https://doi.org/10.35199/dfx2022.09

Krestel, R., Aras, H., Andersson, L., Piroi, F., Hanbury, A., & Alderucci, D. (2021). 2nd Workshop on Patent Text Mining and Semantic Technologies (PatentSemTech2021). *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2693–2696. https://doi.org/10.1145/3404835.3462816

Krestel, R., Aras, H., Andersson, L., Piroi, F., Hanbury, A., & Alderucci, D. (2022). 3rd Workshop on Patent Text Mining and Semantic Technologies (PatentSemTech2022). *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 3474–3477. https://doi.org/10.1145/3477495.3531702

Krestel, R., Aras, H., Andersson, L., Piroi, F., Hanbury, A., & Alderucci, D. (2023). 4th Workshop on Patent Text Mining and Semantic Technologies (PatentSemTech2023). *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 3483–3486. https://doi.org/10.1145/3539618.3591929

Krestel, R., Chikkamath, R., Hewel, C., & Risch, J. (2021). A survey on deep learning for patent analysis. *World Patent Information*, *65*, 102035. https://doi.org/10.1016/j.wpi.2021.102035

Krishna, G. G. (2023). Multilingual NLP. *International Journal of Advanced Engineering and Nano Technology*, *10*(6), 9–12. https://doi.org/10.35940/ijaent.e4119.0610623

Kronemeyer, L. L., Kotzab, H., & Moehrle, M. G. (2022). Analyzing technological competencies in the patent-based supplier portfolio: Introducing an approach for supplier evaluation

using semantic anchor points and similarity measurements. *International Journal of Operations & Production Management*, *42*(11), 1732–1759. https://doi.org/10.1108/IJOPM-09-2021-0607

Lacerda, D. P., Dresch, A., Proença, A., & Antunes Júnior, J. A. V. (2013). Design Science Research: Método de pesquisa para a engenharia de produção. *Gestão & Produção*, *20*(4), 741–761. https://doi.org/10.1590/S0104-530X2013005000014

Lane, P., Koka, B., & Pathak, S. (2006). The Reification of Absorptive Capacity: A Critical Review and Rejuvenation of the Construct. *The Academy of Management Review*, *31*, 833–863. https://doi.org/10.5465/AMR.2006.22527456

Laursen, K., & Salter, A. (2006). Open for Innovation: The Role of Openness in Explaining Innovation Performance Among U.K. Manufacturing Firms. *Strategic Management Journal*, *27*(2), 131–150. https://doi.org/10.1002/smj.507

Lee, C., Kogler, D. F., & Lee, D. (2019). Capturing information on technology convergence, international collaboration, and knowledge flow from patent documents: A case of information and communication technology. *Information Processing & Management*, *56*(4), 1576–1591. https://doi.org/10.1016/j.ipm.2018.09.007

Lee, J., Park, S., & Lee, J. (2022). Study on the Technology Trend Screening Framework Using Unsupervised Learning. *Applied Sciences*, *12*(17), Artigo 17. https://doi.org/10.3390/app12178920

Lee, K.-Y., & Bai, J. (2025). PAI-NET: Retrieval-Augmented Generation Patent Network Using Prior Art Information. *Systems*, *13*(4), Artigo 4. https://doi.org/10.3390/systems13040259

Lee, S., Yoon, B., & Park, Y. (2009). An approach to discovering new technology opportunities: Keyword-based patent map approach. *Technovation*, *29*(6), 481–497. https://doi.org/10.1016/j.technovation.2008.10.006

Lee, Y., Kim, S. Y., Song, I., Park, Y., & Shin, J. (2014). Technology opportunity identification customized to the technological capability of SMEs through two-stage patent analysis. *Scientometrics*, *100*(1), 227–244. https://doi.org/10.1007/s11192-013-1216-0

Lei, L., Qi, J., & Zheng, K. (2019). Patent Analytics Based on Feature Vector Space Model: A Case of IoT. *IEEE Access*, *7*, 45705–45715. IEEE Access. https://doi.org/10.1109/ACCESS.2019.2909123

Levinthal, D., & March, J. (1993). The myopia of learning. *Strategic Management Journal*, *14*, 95–112. https://doi.org/10.1002/SMJ.4250141009

Li, C., Li, W., Hong, Y., & Xiang, H. (2024). A patent retrieval method and system based on double classification. *Information Sciences*, *672*, 120659. https://doi.org/10.1016/j.ins.2024.120659

Li, C., Li, W., Xiang, H., & Hong, Y. (2025). A technical patent map construction method and system based on multi-dimensional technical feature extraction. *Computers in Industry*, *164*, 104167. https://doi.org/10.1016/j.compind.2024.104167

Li, H., Lv, X., & Xu, L. (2017). Automatic Keyword Extraction Method for 3GPP Technical Standard. *2017 International Conference on Computer Systems, Electronics and Control (ICCSEC)*, 762–766. https://doi.org/10.1109/ICCSEC.2017.8446953

Li, R., Wang, X., Liu, Y., & Zhang, S. (2023). Improved Technology Similarity Measurement in the Medical Field based on Subject-Action-Object Semantic Structure: A Case Study of Alzheimer's Disease. *IEEE Transactions on Engineering Management*, *70*(1), 280–293. https://doi.org/10.1109/TEM.2020.3047370

Li, R., Yu, W., Huang, Q., & Liu, Y. (2023). Patent Text Classification based on Deep Learning and Vocabulary Network. *International Journal of Advanced Computer Science and Applications (IJACSA)*, *14*(1), Artigo 1. https://doi.org/10.14569/IJACSA.2023.0140107

Li, S., Hu, J., Cui, Y., & Hu, J. (2018). DeepPatent: Patent classification with convolutional neural networks and word embedding. *Scientometrics*, *117*(2), 721–744. https://doi.org/10.1007/s11192-018-2905-5

Li, X., Wu, Y., Cheng, H., Xie, Q., & Daim, T. (2023). Identifying technology opportunity using SAO semantic mining and outlier detection method: A case of triboelectric nanogenerator technology. *Technological Forecasting and Social Change*, *189*, 122353. https://doi.org/10.1016/j.techfore.2023.122353

Li, Z., & Tate, D. (2010). Patent analysis for systematic innovation: Automatic function interpretation and automatic classification of level of invention using natural language processing and artificial neural networks. *International Journal of Systematic Innovation*, *1*(2), 10–26.

Liang, & Tan. (2007). *A text-mining-based patent analysis in product innovative process*. *250*.

Liang, Tan, & Ma. (2008). *Patent analysis with text mining for TRIZ*. 1147–1151. https://doi.org/10.1109/ICMIT.2008.4654531

Liang, Tan, Wang, & Li. (2009). *Computer-aided classification of patents oriented to TRIZ*. 2389–2393. https://doi.org/10.1109/IEEM.2009.5372983

Liang, Y., Zhou, Y., Zhu, E., & Sun, J. (2024). Entity Identification of Patent Information for Carbon Capture Technology Based on the RoBERTa-BiLSTM-CRF Model. *2024 6th International Conference on Frontier Technologies of Information and Computer (ICFTIC)*, 296–299. https://doi.org/10.1109/ICFTIC64248.2024.10913091

Lim, I. S. S. (2016). The Effectiveness of TRIZ Tools for Eco-Efficient Product Design. Em *Research and Practice on the Theory of Inventive Problem Solving (TRIZ)* (p. 35–54). Springer International Publishing.

Lin, H., Wang, H., Du, D., Wu, H., Chang, B., & Chen, E. (2018). Patent Quality Valuation with Deep Learning Models. Em J. Pei, Y. Manolopoulos, S. Sadiq, & J. Li (Orgs.), *Database Systems for Advanced Applications* (p. 474–490). Springer International Publishing. https://doi.org/10.1007/978-3-319-91458-9_29

Lin, W., Liu, X., & Xiao, R. (2022). Research on Product Core Component Acquisition Based on Patent Semantic Network. *Entropy*, *24*(4), 549. https://doi.org/10.3390/e24040549

Liu, D., Jiang, K., Shalley, C., Keem, S., & Zhou, J. (2016). Motivational mechanisms of employee creativity: A meta-analytic examination and theoretical extension of the creativity literature. *Organizational Behavior and Human Decision Processes*, *137*. https://doi.org/10.1016/j.obhdp.2016.08.001

Liu, Li, Xiong, & Cavallucci. (2020). A new function-based patent knowledge retrieval tool for conceptual design of innovative products. *Computers in Industry*, *115*. https://doi.org/10.1016/j.compind.2019.103154

Liu, W., Tan, R., Li, Z., Cao, G., & Yu, F. (2020). A patent-based method for monitoring the development of technological innovations based on knowledge diffusion. *Journal of Knowledge Management*, *25*(2), 380–401. https://doi.org/10.1108/JKM-09-2019-0502

Liu, W., Zhang, P., & Qiao, W. (2020). Patent Technical Function-effect Representation and Mining Method. *International Conference on Software Engineering and Knowledge Engineering*. https://www.semanticscholar.org/paper/Patent-Technical-Function-effect-Representation-and-Liu-Zhang/4758edd740507dc30c8343de3415be3115e5bd43

Liu, X., Wan, Y., Liu, X., & Zhang, J. (2021). A Patent recommendation algorithm based on topic classification and semantic similarity. *2021 International Conference on Wireless Communications and Smart Grid (ICWCSG)*, 289–292. https://doi.org/10.1109/ICWCSG53609.2021.00063

Liu, Y., Wu, H., Huang, Z., Wang, H., Ma, J., Liu, Q., Chen, E., Tao, H., & Rui, K. (2020). Technical Phrase Extraction for Patent Mining: A Multi-level Approach. *2020 IEEE International Conference on Data Mining (ICDM)*, 1142–1147. https://doi.org/10.1109/ICDM50108.2020.00139

Liu, Y., Wu, H., Huang, Z., Wang, H., Ning, Y., Ma, J., Liu, Q., & Chen, E. (2023). TechPat: Technical Phrase Extraction for Patent Mining. *Transações do ACM na descoberta de conhecimento a partir de dados*, *17*(9), 129:1-129:31. https://doi.org/10.1145/3596603

Liu, Z., Feng, J., & Uden, L. (2023). Technology opportunity analysis using hierarchical semantic networks and dual link prediction. *Technovation*, *128*, 102872. https://doi.org/10.1016/j.technovation.2023.102872

Liwei, Z. (2022). Chinese technical terminology extraction based on DC-value and information entropy. *Scientific Reports*, *12*(1), 20044. https://doi.org/10.1038/s41598-022-23209-6

Lu, H., Liu, X., Yin, Y., & Chen, Z. (2019). A Patent Text Classification Model Based on Multivariate Neural Network Fusion. *2019 6th International Conference on Soft Computing & Machine Intelligence (ISCMI)*, 61–65. https://doi.org/10.1109/ISCMI47871.2019.9004335

Lupu, M. (2017). Information retrieval, machine learning, and Natural Language Processing for intellectual property information. *World Patent Information*, *49*, A1–A3. https://doi.org/10.1016/j.wpi.2017.06.002

Lux, S. (2022). Application of the TRIZ Contradictory Matrix to Foster Innovation for Sustainable Chemical Engineering. *Chemie Ingenieur Technik*, *94*(8), 1071–1079. https://doi.org/10.1002/cite.202100205

Lv, X., Lv, X., You, X., Dong, Z., & Han, J. (2019). Relation Extraction Toward Patent Domain Based on Keyword Strategy and Attention+BiLSTM Model. Em X. Wang, H. Gao, M.

Iqbal, & G. Min (Orgs.), *Collaborative Computing: Networking, Applications and Worksharing* (p. 408–416). Springer International Publishing. https://doi.org/10.1007/978-3-030-30146-0_28

Ma, J.-H., Wang, N.-N., Yao, S., Wei, Z.-M., & Jin, S. (2018). Similar Patent Search Method Based on a Functional Information Fusion. *Proceedings of the 2018 7th International Conference on Software and Computer Applications*, 217–222. https://doi.org/10.1145/3185089.3185130

Ma, T., Zhou, X., Liu, J., Lou, Z., Hua, Z., & Wang, R. (2021). Combining topic modeling and SAO semantic analysis to identify technological opportunities of emerging technologies. *Technological Forecasting and Social Change*, *173*, 121159. https://doi.org/10.1016/j.techfore.2021.121159

Maatens, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, *9*, 2579–2605.

Mafu, M. (2023). Expired patents: An opportunity for higher education institutions. *Frontiers in Research Metrics and Analytics*, *8*. https://www.frontiersin.org/articles/10.3389/frma.2023.1115457

Majewska, O., Razumovskaia, E., Ponti, E. M., Vulić, I., & Korhonen, A. (2023). Cross-Lingual Dialogue Dataset Creation via Outline-Based Generation. *Transactions of the Association for Computational Linguistics*, *11*, 139–156. https://doi.org/10.1162/tacl_a_00539

Makino, K., Sawaguchi, M., & Miyata, N. (2015). Research on Functional Analysis Useful for Utilizing TRIZ. *Procedia Engineering*, *131*, 1021–1030. https://doi.org/10.1016/j.proeng.2015.12.420

Mandl, T. (2009). Artificial Intelligence for Information Retrieval. Em *Encyclopedia of Artificial Intelligence* (p. 151–156). IGI Global Scientific Publishing.

https://www.researchgate.net/publication/314457405_Artificial_Intelligence_for_Information_Retrieval

March, S. T., & Smith, G. F. (1995). Design and natural science research on information technology. *Decision Support Systems, 15*(4), 251–266. https://doi.org/10.1016/0167-9236(94)00041-2

Marmor, A. C., Lawson, W. S., & Terapane, J. F. (1979). The technology assessment and forecast program of the United States patent and trademark office. *World Patent Information, 1*(1), 15–23. https://doi.org/10.1016/0172-2190(79)90006-1

Maskittou, M., Haddadi, A. E., & Routaib, H. (2022). Intelligent technology management based on patent topic modeling. *2022 5th International Conference on Networking, Information Systems and Security: Envisage Intelligent Systems in 5g//6G-based Interconnected Digital Worlds (NISS)*, 1–4. https://doi.org/10.1109/NISS55057.2022.10085417

Masurel, E. (2005). Use of Patent Information: Empirical Evidence from Innovative SMEs. *Research Papers in Economics.* https://www.researchgate.net/publication/4795904_Use_of_Patent_Information_Empirical_Evidence_from_Innovative_SMEs

Mazieri, M. R., Quoniam, L., & Santos, A. M. (2016). Innovation from the patent information: Proposition model Open Source Patent Information Extraction (Crawler). *Revista Gestão & Tecnologia*, *16*(1), 76–112. https://doi.org/10.20397/g&t

McGuinness, D. L., & van Harmelen, F. (2009). *OWL Web Ontology Language Overview*. http://www.w3.org/TR/2004/REC-owl-features-20040210/

McTeague, C., & Chatzimichali, A. (2022). Exploiting patent knowledge in engineering design: A cognitive basis for remodeling patent documents. *Procedia CIRP, 109*, 401–406. https://doi.org/10.1016/j.procir.2022.05.269

Miao, H., Wang, Y., Li, X., & Wu, F. (2022). Integrating Technology-Relationship-Technology Semantic Analysis and Technology Roadmapping Method: A Case of Elderly Smart Wear Technology. *IEEE Transactions on Engineering Management*, *69*(1), 262–278. IEEE Transactions on Engineering Management. https://doi.org/10.1109/TEM.2020.2970972

Miric, M., Jia, N., & Huang, K. G. (2023). Using supervised machine learning for large-scale classification in management research: The case for identifying artificial intelligence patents. *Strategic Management Journal*, *44*(2), 491–519. https://doi.org/10.1002/smj.3441

Montecchi, T., & Russo, D. (2015). *Knowledge based approach for formulating TRIZ contradictions*. *131*, 451–463. https://doi.org/10.1016/j.proeng.2015.12.440

Montgomery, D. P. (2023). This study is not without its limitations: Acknowledging limitations and recommending future research in applied linguistics research articles. *Journal of English for Academic Purposes*, *65*, 101291. https://doi.org/10.1016/j.jeap.2023.101291

Moraes, L. de C., Silvério, I. C., Marques, R. A. S., Anaia, B. de C., Paula, D. F. de, Faria, M. C. S. de, Cleveston, I., Correia, A. de S., & Freitag, R. M. K. (2024). *Análise de ambiguidade linguística em modelos de linguagem de grande escala (LLMs)* (arXiv:2404.16653). arXiv. https://doi.org/10.48550/arXiv.2404.16653

Navas, H. (2013). TRIZ: Design Problem Solving with Systematic Innovation. Em *Advances in Industrial Design Engineering*.

Naveiro, R. M., & de Oliveira, V. M. (2018). QFD and TRIZ integration in product development: A Model for Systematic Optimization of Engineering Requirements. *Production*, *28*. https://doi.org/10.1590/0103-6513.20170093

Ni, X., Samet, A., & Cavallucci, D. (2020). *Build Links Between Problems and Solutions in the Patent*. *597 IFIP*, 64–76. https://doi.org/10.1007/978-3-030-61295-5_6

Ni, X., Samet, A., Chibane, H., & Cavallucci, D. (2021). *PatRIS: Patent Ranking Inventive Solutions*. *12924 LNCS*, 295–309. s. https://doi.org/10.1007/978-3-030-86475-0_29

Nkolongo, F. T., Mehdi, A., & Echchakoui, S. (2024). Application of machine learning in technological forecasting. *Procedia Computer Science*, *251*, 23–30. https://doi.org/10.1016/j.procs.2024.11.080

Noh, H., Jo, Y., & Lee, S. (2015). Keyword selection and processing strategy for applying text mining to patent analysis. *Expert Systems with Applications*, *42*(9), 4348–4360. https://doi.org/10.1016/j.eswa.2015.01.050

Nonaka, I. (1994). A Dynamic Theory of Organizational Knowledge Creation. *Organization Science*, *5*(1), 14–37.

Noor, N., Beram, S., Yuet, F. K. C., Gengatharan, K., & Rasidi, M. S. M. (2023). Bias, Halo Effect and Horn Effect: A Systematic Literature Review. *International Journal of Academic Research in Business and Social Sciences*, *13*(3), 1055–1078.

Nooteboom, B. (2000). *Learning and Innovation in Organizations and Economies*. OUP Oxford.

Oldham, G. R., & Fried, Y. (2016). Job design research and theory: Past, present and future. *Organizational Behavior and Human Decision Processes*, *136*, 20–35. https://doi.org/10.1016/j.obhdp.2016.05.002

O'Leary, D. E. (2013). Artificial Intelligence and Big Data. *IEEE Intelligent Systems*, *28*(2), 96–99. IEEE Intelligent Systems. https://doi.org/10.1109/MIS.2013.39

OpenAI. (2023). *GPT-4 Technical Report*. https://doi.org/arXiv preprint arXiv:2303.08774

Othman, R., Noordin, M. F., Gusmita, R. H., Sembok, T. M. T., & Zulkifli, Z. (2017). *SAO extraction on patent discovery system development for Islamic Finance and Banking*. 59–63. https://doi.org/10.1109/ICT4M.2016.23

Ouzzani, M., Hammady, H., Fedorowicz, Z., & Elmargamid, A. (2016). Rayyan—A web and mobile app for systematic reviews. *Systematic Reviews*, *5*(210). https://doi.org/10.1186/s13643-016-0384-4

Oxford Creativity. (2025a). *About Oxford Creativity*. Oxford Creativity. https://www.triz.co.uk/about-us

Oxford Creativity. (2025b). *Effects database*. Oxford Creativity. http://wbam2244.dns-systems.net/EDB/index.php

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., … Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ (Clinical Research Ed.)*, *372*, n71. https://doi.org/10.1136/bmj.n71

Papagiannopoulou, E., & Tsoumakas, G. (2020). A review of keyphrase extraction. *WIREs Data Mining and Knowledge Discovery*, *10*(2), e1339. https://doi.org/10.1002/widm.1339

Park, H., Kim, K., Choi, S., & Yoon, J. (2013). A patent intelligence system for strategic technology planning. *Expert Systems with Applications*, *40*(7), 2373–2390. https://doi.org/10.1016/j.eswa.2012.10.073

Park, H., Ree, J. J., & Kim, K. (2013). Identification of promising patents for technology transfers using TRIZ evolution trends. *Expert Systems with Applications*, *40*(2), 736–743. https://doi.org/10.1016/j.eswa.2012.08.008

Park, H.-S. (2012). Preliminary Study of Bioinformatics Patents and Their Classifications Registered in the KIPRIS Database. *Genomics & Informatics*, *10*(4), 271–274. https://doi.org/10.5808/GI.2012.10.4.271

Park, S., & Jun, S. (2024). Patent Keyword Analysis Using Regression Modeling Based on Quantile Cumulative Distribution Function. *Electronics*, *13*(21), Artigo 21. https://doi.org/10.3390/electronics13214247

Patton, M. Q. (2015). *Qualitative research & evaluation methods:integrating theory and practice* (4º ed). SAGE Publications.

Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A Design Science Research Methodology for Information Systems Research. *Journal of Management Information Systems*, *24*(3), 45–77. https://doi.org/10.2753/MIS0742-1222240302

Phan, C.-P., Nguyen, H.-Q., & Nguyen, T.-T. (2018). Ontology-based heuristic patent search. *International Journal of Web Information Systems*, *15*. https://doi.org/10.1108/IJWIS-06-2018-0053

Pilkington, A., Lee, L. L., Chan, C. K., & Ramakrishna, S. (2009). Defining key inventors: A comparison of fuel cell and nanotechnology industries. *Technological Forecasting and Social Change*, *76*(1), 118–127. https://doi.org/10.1016/j.techfore.2008.03.015

Pimenta, F. P. (2017). A patente como fonte de informação (des)necessária para a Biotecnologia em Saúde. *Transinformação*, *29*, 323–332. https://doi.org/10.1590/2318-08892017000300009

Pradana, M., Silvianita, A., Madiawati, P. N., Calandra, D., Lanzalonga, F., & Oppioli, M. (2023). A Guidance to Systematic Literature Review to Young Researchers by Telkom University and the University of Turin. *To Maega : Jurnal Pengabdian Masyarakat*, *6*(2), 409–417. https://doi.org/10.35914/tomaega.v6i2.1915

Prickett, P., & Aparicio, I. (2012). The development of a modified TRIZ Technical System ontology. *Computers in Industry*, *63*(3), 252–264. https://doi.org/10.1016/j.compind.2012.01.006

PRISMA. (2025). PRISMA Transparent Reporting of Systematic Reviews and meta-analyses. https://www.prisma-statement.org/prisma-2020-flow-diagram

Pu, K., & Liu, W. (2023). Is absorptive capacity the “panacea” for organizational development? A META analysis of absorptive capacity and firm performance from the perspective of constructivism. *PLOS ONE, 18*(2), e0282321. https://doi.org/10.1371/journal.pone.0282321

Puccetti, G., Chiarello, F., & Fantoni, G. (2021). A simple and fast method for Named Entity context extraction from patents. *Expert Systems with Applications, 184*, 115570. https://doi.org/10.1016/j.eswa.2021.115570

Puccetti, G., Giordano, V., Spada, I., Chiarello, F., & Fantoni, G. (2023). Technology identification from patent texts: A novel named entity recognition method. *Technological Forecasting and Social Change*, *186*, 122160. https://doi.org/10.1016/j.techfore.2022.122160

Quintella, C. M., Meira, M., Guimarães, A. K., Tanajura, A. S., & Da Silva, H. R. G. (2011). Prospecção Tecnológica como uma Ferramenta Aplicada em Ciência e Tecnologia para se Chegar à Inovação. *Revista Virtual de Química*, *3*(5), 406–415.

Ren, Q.-S., Fang, K., Yang, X.-T., & Han, J.-W. (2022). Ensuring the quality of meat in cold chain logistics: A comprehensive review. *Trends in Food Science & Technology*, *119*, 133–151. https://doi.org/10.1016/j.tifs.2021.12.006

Resende Ferreira, V. V., Ricetto, G. C., Gaydeczka, B., Granato, A. C., & Pointer Malpass, G. R. (2022). Patents, what are they good for? Academic chemistry researcher’s perceptions of patents and their importance. *World Patent Information*, *70*, 102124. https://doi.org/10.1016/j.wpi.2022.102124

Reymond, D., & Quoniam, L. (2016). A new patent processing suite for academic and research purposes. *World Patent Information, 47,* 40–50. https://doi.org/10.1016/j.wpi.2016.10.001

Rivera-Garrido, N., Ramos-Sosa, M. P., Accerenzi, M., & Brañas-Garza, P. (2022). Continuous and binary sets of responses differ in the field. *Scientific Reports, 12*, 14376. https://doi.org/10.1038/s41598-022-17907-4

Robert, T., & Mayer, F. (2003). Improving Models for Better Knowledge Interoperability in Product Design Process. *IFAC Proceedings Volumes*, *36*(22), 233–237. https://doi.org/10.1016/S1474-6670(17)37723-6

Rogers, M., Helmers, C., Hall, B. H., & Sena, V. (2012, novembro 16). *The Use of Alternatives to Patents and Limits to Incentives*. https://papers.ssrn.com/abstract=2710628

Rose, S., Engel, D., Cramer, N., & Cowley, W. (2010). Automatic Keyword Extraction from Individual Documents. Em *Text Mining* (p. 1–20). John Wiley & Sons, Ltd. https://doi.org/10.1002/9780470689646.ch1

Rosenberg, C., Hebert, M., & Schneiderman, H. (2005). Semi-Supervised Self-Training of Object Detection Models. *2005 Seventh IEEE Workshops on Applications of Computer Vision (WACV/MOTION'05) - Volume 1*, *1*, 29–36. https://doi.org/10.1109/ACVMOT.2005.107

Rossi, J., Wirth, M., & Kanoulas, E. (2019). *Query Generation for Patent Retrieval with Keyword Extraction based on Syntactic Features*. International Conference on Legal Knowledge and Information Systems. https://doi.org/10.48550/arXiv.1906.07591

Rüdiger, M., Antons, D., & Salge, T. O. (2017). From Text to Data: On The Role and Effect of Text Pre-Processing in Text Mining Research. *Academy of Management Proceedings*, *2017*(1), 16353. https://doi.org/10.5465/AMBPP.2017.16353abstract

Russo, D. (2011). *Knowledge extraction from patent: Achievements and open problems. A multidisciplinary approach to find functions*. 567–576. https://doi.org/10.1007/978-3-642-15973-2_57

Russo, D., Carrara, P., & Facoetti, G. (2018). Technical problem identification for supervised state of the art. *IFAC-PapersOnLine*, *51*(11), 1341–1346. https://doi.org/10.1016/j.ifacol.2018.08.344

Russo, D., & Gervasoni, D. (2022). *AI Based Patent Analyzer for Suggesting Solutive Actions and Graphical Triggers During Problem Solving*. *655 IFIP*, 187–197. https://doi.org/10.1007/978-3-031-17288-5_17

Russo, D., & Montecchi, T. (2011). *A function-behaviour oriented search for patent digging*. *2*(PARTS A AND B), 1111–1120. https://doi.org/10.1115/DETC2011-47733

Russo, D., Montecchi, T., & Ying, L. (2012). *Functional-based search for patent technology transfer*. *2*, 529–539. https://doi.org/10.1115/DETC2012-70833

Russo D., S. C., Carrara P. (2020). *How to Organize a Knowledge Basis Using TRIZ Evolution Tree: A Case About Sustainable Food Packaging* (rayyan-1180749954). *597*.

Ryu, S., & Lee, S. (2024). Development of a technology tree using patent information. *Advanced Engineering Informatics*, *59*, 102277. https://doi.org/10.1016/j.aei.2023.102277

Saad, F. (2019). Named Entity Recognition for Biomedical Patent Text using Bi-LSTM Variants. *Proceedings of the 21st International Conference on Information Integration and Web-Based Applications & Services*, 617–621. https://doi.org/10.1145/3366030.3366104

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, *24*(5), 513–523. https://doi.org/10.1016/0306-4573(88)90021-0

Sanches, C., Meireles, M., & Silva, O. (2014). Framework for the generic process of diagnosis in quality problem solving. *Total Quality Management & Business Excellence*, *26*, 1–15. https://doi.org/10.1080/14783363.2014.918707

Sarica, S., Luo, J., & Wood, K. L. (2020). TechNet: Technology semantic network based on patent data. *Expert Systems with Applications*, *142*, 112995. https://doi.org/10.1016/j.eswa.2019.112995

Sarica, S., Song, B., Low, E., & Luo, J. (2019). Engineering Knowledge Graph for Keyword Discovery in Patent Search. *Proceedings of the Design Society: International Conference on Engineering Design*, 2249–2258. https://doi.org/10.1017/dsi.2019.231

Savransky, S. D. (2000). *Engineering of creativity (Introduction to TRIZ Methodology of Inventive Problem Solving*. CRC Press.

Shaheen, N., Shaheen, A., Ramadan, A., Hefnawy, M., Ramadan, A., A. Ibrahim, I., Hassanein, M., Ashour, M., & Flouty, O. (2023). Appraising Systematic Reviews: A Comprehensive Guide to Ensuring Validity and Reliability. *Frontiers in Research Metrics and Analytics*, *8*. https://doi.org/10.3389/frma.2023.1268045

Shalaby, W., & Zadrozny, W. (2019). Patent retrieval: A literature review. *Knowledge and Information Systems*, *61*(2), 631–660. https://doi.org/10.1007/s10115-018-1322-7

Shelick, E. G. G. (2009). *Variácion Denominativa y conceptual de la terminología de textos de patentes en el ámbito mexicano* [Dissertação de Mestrado, Universidad Nacional Autónoma de México]. https://repositorio.unam.mx/contenidos/216103

Shi, X., Feng, Z., Liu, J., Cheng, Q., & Lu, W. (2022). *Automatic Construction of Technology Function Matrix*. *3210*. https://ceur-ws.org/Vol-3210/paper9.pdf

Shin, H., Lee, H. J., & Cho, S. (2023). General-use unsupervised keyword extraction model for keyword analysis. *Expert Systems with Applications*, *233*, 120889. https://doi.org/10.1016/j.eswa.2023.120889

Siddharth, L., Blessing, L. T. M., Wood, K. L., & Luo, J. (2022). Engineering Knowledge Graph From Patent Database. *Journal of Computing and Information Science in Engineering*, *22*(2). https://doi.org/10.1115/1.4052293

Simon, H. A. (1996). *The Sciences of the Artificial* (3º ed). The MIT Press.

Singh. (2018). *Natural Language Processing for Information Extraction*. https://doi.org/10.48550/arXiv.1807.02383

Son, J., Moon, H., Lee, J., Lee, S., Park, C., Jung, W., & Lim, H. (2022). AI for Patents: A Novel Yet Effective and Efficient Framework for Patent Analysis. *IEEE Access*, *10*, 59205–59218. IEEE Access. https://doi.org/10.1109/ACCESS.2022.3176877

Soo, V.-W., Lin, S.-Y., Yang, S.-Y., Lin, S.-N., & Cheng, S.-L. (2005). A cooperative multi-agent platform for invention based on ontology and patent document analysis. *Proceedings of the Ninth International Conference on Computer Supported Cooperative Work in Design, 2005.*, *1*, 411-416 Vol. 1. https://doi.org/10.1109/CSCWD.2005.194207

Souchkov, V. (2016). A Glossary of Essential TRIZ Terms. Em *Research and Practice on the Theory of Inventive Problem Solving (TRIZ): Linking Creativity, Engineering and Innovation* (p. 265–281). Springer.

Souchkov, V. (2017). *Accelerate Innovation with TRIZ*. https://www.researchgate.net/publication/332946994_Accelerate_Innovation_with_TRIZ#fullTextFileContent

Souili, A., & Cavallucci, D. (2013). *Toward an automatic extraction of idm concepts from patents*. 115–124. https://doi.org/10.1007/978-1-4471-4507-3_12

Souili, A., Cavallucci, D., & Rousselot, F. (2015a). *Identifying and reformulating knowledge items to fit with the Inventive Design Method (IDM) model for a semantically-based patent mining*. *131*, 1130–1139. https://doi.org/10.1016/j.proeng.2015.12.432

Souili, A., Cavallucci, D., & Rousselot, F. (2015b). *Natural Language Processing (NLP)—A solution for knowledge extraction from patent unstructured data*. *131*, 635–643. https://doi.org/10.1016/j.proeng.2015.12.457

Souili, A., Cavallucci, D., Rousselot, F., & Zanni. (2015). *Starting from patents to find inputs to the Problem Graph model of IDM-TRIZ*. *131*, 150–161. https://doi.org/10.1016/j.proeng.2015.12.365

Souili, Cavallucci, & Rousselot. (2015c). *A lexico-syntactic pattern matching method to extract IDM- TRIZ knowledge from on-line patent databases*. *131*, 418–425. https://doi.org/10.1016/j.proeng.2015.12.437

Spreafico, C., & Spreafico, M. (2021). Using text mining to retrieve information about circular economy. *Computers in Industry*, *132*. https://doi.org/10.1016/j.compind.2021.103525

Srinivasan, R., & Kraslawski, A. (2006). Application of the TRIZ creativity enhancement approach to design of inherently safer chemical processes. *Chemical Engineering and Processing: Process Intensification*, *45*(6), 507–514. https://doi.org/10.1016/j.cep.2005.11.009

Stamatis, V., Salampasis, M., & Diamantaras, K. (2024). A novel re-ranking architecture for patent search. *World Patent Information*, *78*, 102282. https://doi.org/10.1016/j.wpi.2024.102282

Stettler, T. R., Moosauer, E. J., Schweiger, S. A., Baldauf, A., & Audretsch, D. (2024). Absorptive capacity in a more (or less) absorptive environment: A meta-analysis of

contextual effects on firm innovation. *Journal of Product Innovation Management*, *42*(1), 18–47. https://doi.org/10.1111/jpim.12758

Stoian, M.-C., Tardios, J. A., & Samdanis, M. (2024). The knowledge-based view in international business: A systematic review of the literature and future research directions. *International Business Review*, *33*(2), 102239. https://doi.org/10.1016/j.ibusrev.2023.102239

Suárez-Figueroa, M. C., Gómez-Pérez, A., & Villazón-Terrazas, B. (2009). *How to Write and Use the Ontology Requirements Specification Document* (R. Meersman, T. Dillon, & P. Herrero, Orgs.). Springer.

Sun, Y., Liu, W., Cao, G., Peng, Q., Gu, J., & Fu, J. (2022). Effective design knowledge abstraction from Chinese patents based on a meta-model of the patent design knowledge graph. *Computers in Industry*, *142*, 103749. https://doi.org/10.1016/j.compind.2022.103749

Sun, Y.-D., Cao, G.-Z., Gao, C., Yang, W.-D., Han, W.-P., & Wang, K. (2021). Extraction and Modeling of Chinese Patent Information for Technical Advancement Evaluation. *IFIP Advances in Information and Communication Technology*, *635 IFIP*, 127–140. https://doi.org/10.1007/978-3-030-86614-3_10

Suominen, H., Ferraro, G., Nualart Vilaplana, J., & Hanlen, L. (2018, fevereiro 1). *User Study for Measuring Linguistic Complexity and Its Reduction by Technology on a Patent Website*. 34 International Conference on Machine Learning, Sydney, Australia.

Taduri, S., Law, K. H., Kesan, J. P., & Sriram, R. D. (2019). Utilization of Bio-Ontologies for Enhancing Patent Information Retrieval. *2019 IEEE 43rd Annual Computer Software and Applications Conference (COMPSAC)*, *2*, 91–96. https://doi.org/10.1109/COMPSAC.2019.10189

Takeuchi, H. (2013). Knowledge-Based View of Strategy. *Universia Business Review*, *40*, 68–79.

Teece, D. J. (1986). Profiting from technological innovation: Implications for integration, collaboration, licensing and public policy. *Research Policy*, *15*(6), 285–305. https://doi.org/10.1016/0048-7333(86)90027-2

Teece, D. J., Pisano, G., & Shuen, A. (1997). Dynamic capabilities and strategic management. *Strategic Management Journal*, *18*(7), 509–533. https://doi.org/10.1002/(SICI)1097-0266(199708)18:7<509::AID-SMJ882>3.0.CO;2-Z

Teng, H., Wang, N., Zhao, H., Hu, Y., & Jin, H. (2024). Enhancing semantic text similarity with functional semantic knowledge (FOP) in patents. *Journal of Informetrics*, *18*(1), 101467. https://doi.org/10.1016/j.joi.2023.101467

Tian, C., Zhang, J., Liu, D., Wang, Q., & Lin, S. (2022). Technological topic analysis of standard-essential patents based on the improved Latent Dirichlet Allocation (LDA) model. *Technology Analysis & Strategic Management*, *36*(9), 2084–2099. https://doi.org/10.1080/09537325.2022.2130039

Tian, C., Zhang ,Junyan, Liu ,Dayong, Wang ,Qing, & and Lin, S. (2024). Technological topic analysis of standard-essential patents based on the improved Latent Dirichlet Allocation (LDA) model. *Technology Analysis & Strategic Management*, *36*(9), 2084–2099. https://doi.org/10.1080/09537325.2022.2130039

Trapp, S., Großer, N., & Warschat, J. (2023). *Question Answering with Transformers and Few-Shot Learning to Find Inventive Solutions for IDM-TRIZ Problems and Contradictions in Patents*. *682*, 23–42. https://doi.org/10.1007/978-3-031-42532-5_2

Trapp, S., & Warschat, J. (2025). LLM-Based Extraction of Contradictions from Patents. *IFIP Advances in Information and Communication Technology*, *735 IFIP*, 3–19. https://doi.org/10.1007/978-3-031-75919-2_1

Trappey, A. J. C., Wei, A. Y. E., Chen, N. K. T., Li, K.-A., Hung, L. P., & Trappey, C. V. (2023). Patent landscape and key technology interaction roadmap using graph convolutional network – Case of mobile communication technologies beyond 5G. *Journal of Informetrics*, *17*(1). https://ideas.repec.org//a/eee/infome/v17y2023i1s1751157722001079.html

Trappey, A., Lin, G.-B., & Hung, L.-P. (2024). Intelligent Text Mining for Ontological Knowledge Graph Refinement and Patent Portfolio Analysis—Case Study of Net-Zero Data Center Innovation Management. *Information*, *15*(7), 374. https://doi.org/10.3390/info15070374

Trappey, A., Trappey, C., & Chang, A.-C. (2020). Intelligent Extraction of a Knowledge Ontology From Global Patents: The Case of Smart Retailing Technology Mining. *International Journal on Semantic Web and Information Systems*, *16*, 61–80. https://doi.org/10.4018/IJSWIS.2020100104

Trappey, A., Trappey, C., Wu, & Wang. (2020). Intelligent compilation of patent summaries using machine learning and natural language processing techniques. *Advanced Engineering Informatics*, *43*. https://doi.org/10.1016/j.aei.2019.101027

Trappey, Trappey, C. V., Chen, C. H., & Anggrahini, D. (2024). Ontology-based patent analysis for bike-sharing services: Identifying competitive advantages of a product-service business. *Asia Pacific Management Review*. https://doi.org/10.1016/j.apmrv.2024.09.005

Trappey, Trappey, Wu, Liaw, & Zhang. (2013). *Development of innovative product design process using patent multi-scale analysis and TRIZ methodology*. *2*, 1215–1225. https://www.scopus.com/inward/record.uri?eid=2-s2.0-84898770389&partnerID=40&md5=55f4df7952c8eeab0c24721fbe4e172b

Trappey, Trappey, & Zou. (2018). Using non-supervised machine learning to generate a knowledge ontology for patent analytics. *Proceedings of International Conference on Computers and Industrial Engineering*. 48th International Conference on Computers and Industrial Engineering, CIE 2018, Auckland, New Zealand. https://www.semanticscholar.org/paper/Using-non-supervised-machine-learning-to-generate-a-Trappey-Trappey/96c9ade3092e8481265dc5dc91ad92587b4dcc8b

Triguero, I., García, S., & Herrera, F. (2015). Self-labeled techniques for semi-supervised learning: Taxonomy, software and empirical study. *Conhecimento e Sistemas de Informação*, *42*(2), 245–284. https://doi.org/10.1007/s10115-013-0706-y

Trippe, A. (2015). *Guidelines prepared for the patent landscape reports*. World Intellectual Property Organization. https://www.cambodiaip.gov.kh/DocResources/b902659f-bc5c-42e2-8c9e-ffe07d34342e_2f1a3b22-ccf9-4e33-9247-94627e014e79-en.pdf

Tseng, Y.-H., Lin, C.-J., & Lin, Y.-I. (2007). Text mining techniques for patent analysis. *Information Processing & Management*, *43*(5), 1216–1247. https://doi.org/10.1016/j.ipm.2006.11.011

Uzzi, B. (1997). Social Structure and Competition in Interfirm Networks: The Paradox of Embeddedness. *Administrative Science Quarterly*, *42*(1), 35–67. https://doi.org/10.2307/2393808

Valverde, U. Y., Nadeau, J.-P., & Scaravetti, D. (2017). A new method for extracting knowledge from patents to inspire designers during the problem-solving phase. *Journal of Engineering Design*, *28*(6), 369–407. https://doi.org/10.1080/09544828.2017.1316361

van Aken, J. E. (2004). Management Research Based on the Paradigm of the Design Sciences: The Quest for Field-Tested and Grounded Technological Rules. *Journal of Management Studies*, *41*(2), 219–246. https://doi.org/10.1111/j.1467-6486.2004.00430.x

Venable, J. (2006). The role of theory and theorising in design science research. *First International Conference on Design Science Research in Information Systems and Technology*.

Vereschak, G., & Korobkin, D. (2019). *Identification of Descriptions of Scientific-Technical Effects in Patent Documents*. 24–27.

Verhaegen, P.-A., D'hondt, J., Vertommen, J., Dewulf, S., & Duflou, J. R. (2011). *Searching for similar products through patent analysis*. *9*, 431–441. https://doi.org/10.1016/j.proeng.2011.03.131

Verhaegen, P.-A., D'hondt, J., Vertommen, J., Dewulf, S., & Duflou, J. R. (2014). *Interrelating products through properties via patent analysis*. 252–257. https://www.scopus.com/inward/record.uri?eid=2-s2.0-84912074178&partnerID=40&md5=135b6c305c379f80a33fb3b6b5c44c6e

Vicente Gomila, J. M., & Palop Marro, F. (2013). Combining tech-mining and semantic-TRIZ for a faster and better technology analysis: A case in energy storage systems. *Technology Analysis and Strategic Management*, *25*(6), 725–743. https://doi.org/10.1080/09537325.2013.803065

Vicente-Gomila, J. M. (2014). The contribution of syntactic-semantic approach to the search for complementary literatures for scientific or technical discovery. *Scientometrics*, *100*(3), 659–673. https://doi.org/10.1007/s11192-014-1299-2

Vicente-Gomila, J. M., Palli, A., de la Calle, B., Artacho, M. A., & Jimenez, S. (2017). Discovering shifts in competitive strategies in probiotics, accelerated with TechMining. *Scientometrics*, *111*(3), 1907–1923. https://doi.org/10.1007/s11192-017-2339-5

Viglioni, M. T. D., Calegario, C. L. L., Aveline, C. E. S., Ferreira, M. P., Borini, F. M., & Bruhn, N. C. P. (2023). Effects of intellectual property rights on innovation and economic

activity: A non-linear perspective from Latin America. *Structural Change and Economic Dynamics*, *67*, 359–371. https://doi.org/10.1016/j.strueco.2023.09.001

Vincent, J., & Cavallucci, D. (2018). Development of an Ontology of Biomimetics Based on Altshuller's Matrix. Em D. Cavallucci, R. De Guio, & S. Koziołek (Orgs.), *Automated Invention for Smart Industries* (p. 14–25). Springer International Publishing. https://doi.org/10.1007/978-3-030-02456-7_2

Wang, Chang, & Kao. (2010). Identifying technology trends for R and D planning using TRIZ and text mining. *R and D Management*, *40*(5), 491–509. https://doi.org/10.1111/j.1467-9310.2010.00612.x

Wang, F., Tan, R., Wang, K., Cen, S., & Peng, Q. (2024). Innovative product design based on radical problem solving. *Computers & Industrial Engineering*, *189*, 109941. https://doi.org/10.1016/j.cie.2024.109941

Wang, J., & Chen, Y.-J. (2019). A novelty detection patent mining approach for analyzing technological opportunities. *Advanced Engineering Informatics*, *42*, 100941. https://doi.org/10.1016/j.aei.2019.100941

Wang, J., Ding, Z., Liu, Z., & Feng, L. (2024). Technology opportunity discovery based on patent analysis: A hybrid approach of subject-action-object and generative topographic mapping. *Technology Analysis & Strategic Management*. https://www.tandfonline.com/doi/abs/10.1080/09537325.2022.2126306

Wang, J., Omar, A. H., Alotaibi, F. M., Daradkeh, Y. I., & Althubiti, S. A. (2022). Business intelligence ability to enhance organizational performance and performance evaluation capabilities by improving data mining systems for competitive advantage. *Information Processing & Management*, *59*(6), 103075. https://doi.org/10.1016/j.ipm.2022.103075

Wang, J., Song, F., Walia, K., Farber, J., & Dara, R. (2019). Using Convolutional Neural Networks to Extract Keywords and Keyphrases: A Case Study for Foodborne Illnesses. *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*, 1398–1403. https://doi.org/10.1109/ICMLA.2019.00228

Wang, M., Hu, X., Xie, P., & Du, Y. (2021). *Automatic Construction of a Domain-specific Knowledge Graph for Chinese Patent Based on Information Extraction*. 1–8. https://doi.org/10.1109/ICMSSE53595.2021.00008

Wang, T., Zhao ,Yushan, Zhu ,Guangli, Liu ,Yunduo, Li ,Hanchen, Zhang ,Shunxiang, & and Hsieh, M. (2024). An entity and relation extraction model based on context query and axial attention towards patent texts. *Connection Science*, *36*(1), 2426816. https://doi.org/10.1080/09540091.2024.2426816

Wang, Tian, Geng, Evans, & Che. (2016). *Extraction of Principle Knowledge from Process Patents for Manufacturing Process Innovation*. *56*, 193–198. https://doi.org/10.1016/j.procir.2016.10.053

Wang, Z., Guo, W., Shao, H., Wang, L., Chang, Z., Zhang, Y., & Liu, Z. (2024). From technology opportunities to solutions generation via patent analysis: Application of machine learning-based link prediction. *Advanced Engineering Informatics*, *62*, 102944. https://doi.org/10.1016/j.aei.2024.102944

Wang, Z., & Liu, Y. (2022). SEA-PS: Semantic embedding with attention to measuring patent similarity by leveraging various text fields. *Journal of Information Science*, 01655515221106651. https://doi.org/10.1177/01655515221106651

Wang, Z., & Liu, Y. (2024). SEA-PS: Semantic embedding with attention to measuring patent similarity by leveraging various text fields. *Journal of Information Science*, *50*(4), 831–850. https://doi.org/10.1177/01655515221106651

Watanabe, K., & Zhou, Y. (2022). Theory-Driven Analysis of Large Corpora: Semisupervised Topic Classification of the UN Speeches. *Social Science Computer Review*, *40*(2), 346–366. https://doi.org/10.1177/0894439320907027

Wei, T., Jiang, T., Feng, D., & Xiong, J. (2023). Exploring the Evolution of Core Technologies in Agricultural Machinery: A Patent-Based Semantic Mining Analysis. *Electronics*, *12*(20), 4277. https://doi.org/10.3390/electronics12204277

Wickert, C., Post, C., Doh, J. P., Prescott, J. E., & Prencipe, A. (2021). Management Research that Makes a Difference: Broadening the Meaning of Impact. *Journal of Management Studies*, *58*(2), 297–320. https://doi.org/10.1111/joms.12666

World Intellectual Property Organization. (2024). *Global innovation index 2024: Unlocking the Promise of Social Entrepreneurship*. World Intellectual Property Organization. https://www.wipo.int/web-publications/global-innovation-index-2024/en/

World Intellectual Property Organization. (2025a). *International Patent Classification (IPC)*. Classification IPC. https://www.wipo.int/web/classification-ipc

World Intellectual Property Organization. (2025b). *Search International and National Patent Collections*. https://patentscope.wipo.int/search/en/search.jsf

Worren, N., Moore, K., & Elliott, R. (2002). When Theories Become Tools: Toward a Framework for Pragmatic Validity. *Human Relations - HUM RELAT*, *55*, 1227–1250. https://doi.org/10.1177/0018726702055010082

Wu, H., Yang, H., Ma, J., & Tan, R. (2010). Function-based patent retrieval for concept design. *2010 IEEE International Conference on Industrial Engineering and Engineering Management*, 348–351. https://doi.org/10.1109/IEEM.2010.5674304

Wu, J.-L. (2019). Patent Quality Classification System Using the Feature Extractor of Deep Recurrent Neural Network. *2019 IEEE International Conference on Big Data and Smart*

Computing (BigComp), 1–8. 2019 IEEE International Conference on Big Data and Smart Computing (BigComp). https://doi.org/10.1109/BIGCOMP.2019.8679141

Xiao, L., Wang, G., & Zuo, Y. (2018). Research on Patent Text Classification Based on Word2Vec and LSTM. *2018 11th International Symposium on Computational Intelligence and Design (ISCID)*, *01*, 71–74. https://doi.org/10.1109/ISCID.2018.00023

Xie, Z., & Miyazaki, K. (2013). Evaluating the effectiveness of keyword search strategy for patent identification. *World Patent Information*, *35*(1), 20–30. https://doi.org/10.1016/j.wpi.2012.10.005

Xiong, F., Li, S., Zhang, W., & Xu, N. (2024). Individuals' capacity to innovate: A literature review of individual absorptive capacity. *Innovation*, *27*(4), 564–584. https://doi.org/10.1080/14479338.2024.2363255

Xu, J., Guo, L., Jiang, J., Ge, B., & Li, M. (2019). A deep learning methodology for automatic extraction and discovery of technical intelligence. *Technological Forecasting and Social Change*, *146*, 339–351. https://doi.org/10.1016/j.techfore.2019.06.004

Xu, Y., Hu, L., Zhao, J., Qiu, Z., XU, K., Ye, Y., & Gu, H. (2025). A Survey on Multilingual Large Language Models: Corpora, Alignment, and Bias. *Frontiers of Computer Science*, *19*(11). https://doi.org/10.1007/s11704-024-40579-4

Xu, Y., Liu, Y., & Li, Z. (2019). How Different Scientific Cultures Influence Triz Innovations: Applying Actor–Network Theory in Case Studies of Tesla and NIO Electric Cars. *Cultures of Science*, *2*(2), 81–96. https://doi.org/10.1177/209660831900200202

Xu, Yue, Z., Pang, Elahi, Li, & Wang. (2022). Integrative model for discovering linked topics in science and technology. *Journal of Informetrics*, *16*(2), 101265. https://doi.org/10.1016/j.joi.2022.101265

Yan, W., Liu, H., Zanni-Merk, C., & Cavallucci, D. (2015). IngeniousTRIZ: An automatic ontology-based system for solving inventive problems. *Knowledge-Based Systems*, *75*, 52–65. https://doi.org/10.1016/j.knosys.2014.11.015

Yang, Y., Wu, Z., Yang, Y., Lian, S., Guo, F., & Wang, Z. (2022). A Survey of Information Extraction Based on Deep Learning. *Applied Sciences*, *12*(19), Artigo 19. https://doi.org/10.3390/app12199691

Yoon, B. (2010). Strategic visualisation tools for managing technological information. *Technology Analysis & Strategic Management*, *22*(3), 377–397. https://doi.org/10.1080/09537321003647438

Yoon, B., Kim, S., Kim, S., & Seol, H. (2022). Doc2vec-based link prediction approach using SAO structures: Application to patent network. *Scientometrics*, *127*(9), 5385–5414. https://doi.org/10.1007/s11192-021-04187-4

Yoon, B., & Magee, C. L. (2018). Exploring technology opportunities by visualizing patent information based on generative topographic mapping and link prediction. *Technological Forecasting and Social Change*, *132*, 105–117. https://doi.org/10.1016/j.techfore.2018.01.019

Yoon, J., & Kim, K. (2012). TrendPerceptor: A property-function based technology intelligence system for identifying technology trends from patents. *Expert Systems with Applications*, *39*(3), 2927–2938. https://doi.org/10.1016/j.eswa.2011.08.154

Yoon, J., Park, H., & Kim, K. (2013). Identifying technological competition trends for R&D planning using dynamic patent maps: SAO-based content analysis. *Scientometrics*, *94*(1), 313–331. https://doi.org/10.1007/s11192-012-0830-6

Yoon, & Park. (2004). A text-mining-based patent network: Analytical tool for high-technology trend. *The Journal of High Technology Management Research*, *15*(1), 37–50. https://doi.org/10.1016/j.hitech.2003.09.003

Yoon, & Park. (2005). A systematic approach for identifying technology opportunities: Keyword-based morphology analysis. *Technological Forecasting and Social Change*, *72*(2), 145–160. https://doi.org/10.1016/j.techfore.2004.08.011

Yue, G., Liu, J., Hou, Y., & Zhang, Q. (2023). A Novel Patent Knowledge Extraction Method for Innovative Design. *IEEE Access*, *11*, 2182–2198. IEEE Access. https://doi.org/10.1109/ACCESS.2022.3229490

Yue, G., Liu, J., Zhang, Q., & Hou, Y. (2023). Building a Design-Rationale-Centric Knowledge Network to Realize the Internalization of Explicit Knowledge. *Applied Sciences*, *13*(3), Artigo 3. https://doi.org/10.3390/app13031539

Yun, S., Cho, W., Kim, C., & Lee, S. (2022). Technological trend mining: Identifying new technology opportunities using patent semantic analysis. *Information Processing & Management*, *59*(4), 102993. https://doi.org/10.1016/j.ipm.2022.102993

Zahra, & George. (2002). Absorptive Capacity: A Review, Reconceptualization, and Extension. *Academy of Management Review*, *27*(2). https://doi.org/10.2307/4134351

Zanella, G., Liu, C. Z., & Choo, K.-K. R. (2023). Understanding the Trends in Blockchain Domain Through an Unsupervised Systematic Patent Analysis. *IEEE Transactions on Engineering Management*, *70*(6), 1991–2005. https://doi.org/10.1109/TEM.2021.3074310

Zaniro, D. L., Quoniam, L., de Souza, M. G., & Segundo, W. L. R. de C. (2024). *Towards an open TRIZ Multilingual database*. The 19th International Conference on Open Repositories, Göteborg, Sweden.

Zhai, D., Li, M., & Cai, W. (2020). *TRIZ Technical Contradiction Extraction Method Based on Patent Semantic Space Mapping*. 125–130. https://doi.org/10.1145/3414752.3414802

Zhang, C., Jian, S., Chao, L., Fan, C., Bo, L., & Chen, L. (2023). A Semantic Understanding Method for Patent Text Based on Large Language Model. *2023 IEEE 7th Conference on Energy Internet and Energy System Integration (EI2)*, 2179–2182. https://doi.org/10.1109/EI259745.2023.10513309

Zhang, J., Li, K., & Yao, C. (2018). Event-based Summarization for Scientific Literature in Chinese. *Procedia Computer Science*, *129*, 88–92. https://doi.org/10.1016/j.procs.2018.03.052

Zhang, J., Liu, Y., Jiang, L., & Shi, J. (2022). Discovery of topic evolution path and semantic relationship based on patent entity representation. *Aslib Journal of Information Management*, *75*(3), 618–642. https://doi.org/10.1108/AJIM-03-2022-0124

Zhang, J., & Yu, W. (2020). Early detection of technology opportunity based on analogy design and phrase semantic representation. *Scientometrics*, *125*(1), 551–576. https://doi.org/10.1007/s11192-020-03641-z

Zhang, L., Liu, Z., Li, L., Shen, C., & Li, T. (2018). PatSearch: An integrated framework for patentability retrieval. *Knowledge and Information Systems*, *57*. https://doi.org/10.1007/s10115-017-1127-0

Zhang, L., Sun, X., Ma, X., & Hu, K. (2024). A New Entity Relationship Extraction Method for Semi-Structured Patent Documents. *Electronics*, *13*(16), 3144. https://doi.org/10.3390/electronics13163144

Zhang, L., Zhao, J., Lu, H., Gong, L., Li, L., Zheng, J., Li, H., & Zhu, Z. (2011). High sensitive and selective formaldehyde sensors based on nanoparticle-assembled ZnO micro-

octahedrons synthesized by homogeneous precipitation method. *Sensors and Actuators B: Chemical*, *160*(1), 364–370. https://doi.org/10.1016/j.snb.2011.07.062

Zhang, Tan, Peng, Shao, Dong, & Wang. (2022). Construction and Application of Enterprise Knowledge Base for Product Innovation Design. *Applied Sciences (Switzerland)*, *12*(13). https://doi.org/10.3390/app12136358

Zhang, Wang, & Nie. (2022). Agile innovation process model based on computer-aided patent knowledge mining and functional analogy. *Computer-Aided Design and Applications*, *19*(2), 346–374. https://doi.org/10.14733/CADAPS.2022.346-374

Zhang, Wu, Liu, Qin, & Zhou. (2023). Identification of Product Innovation Path Incorporating the FOS and BERTopic Model from the Perspective of Invalid Patents. *Applied Sciences (Switzerland)*, *13*(13). https://doi.org/10.3390/app13137987

Zhang, Y., Li, S., Chen, X., Qian, F., Zhao, S., Zhu, S., & Wang, Y. (2020). Semantic Based Heterogeneous Information Network Embedding for Patent Citation Recommendation. *2020 International Conference on Artificial Intelligence and Computer Engineering (ICAICE)*, 518–527. https://doi.org/10.1109/ICAICE51518.2020.00106

Zhao, X. (2024). Big Data Related Patent Retrieval System Based on Filtering Rules. *Journal of Computing and Information Technology*, *32*(4), 265–276. https://doi.org/10.20532/cit.2024.1005865

Zheng, Q., Guo ,Kefu, & and Xu, L. (2024). A large-scale Chinese patent dataset for information extraction. *Systems Science & Control Engineering*, *12*(1), 2365328. https://doi.org/10.1080/21642583.2024.2365328

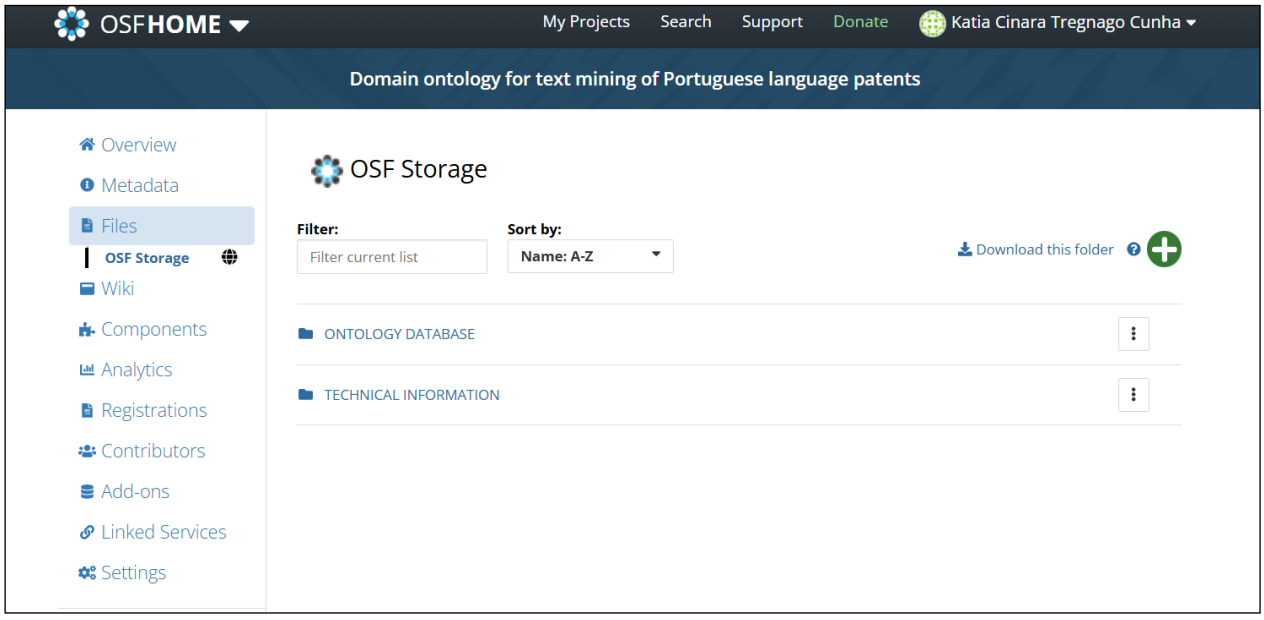
Zheng, S., Zhang, W., & Du, J. (2011). Knowledge-based dynamic capabilities and innovation in networked environments. *Journal of Knowledge Management*, *15*(6), 1035–1051. https://doi.org/10.1108/13673271111179352

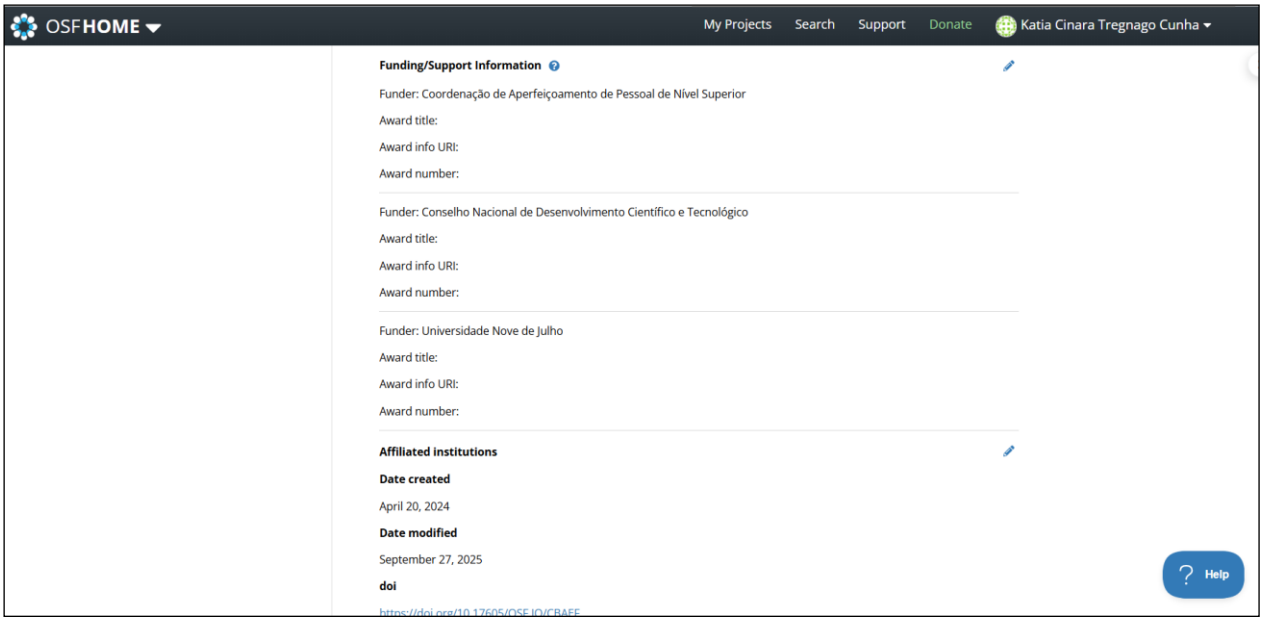
Zheng, W., Ding, Q., Li, N., Pan, Y., & Dong, X. (2024). Patent Recommendation Methods Based on Transformer Encoders and Regularization Strategies. *2024 IEEE International Conference on Advanced Information, Mechanical Engineering, Robotics and Automation (AIMERA)*, 1–8. https://doi.org/10.1109/AIMERA59657.2024.10735761

Zhou, P., Jiang, X., & Zhao, S. (2024). Unsupervised technical phrase extraction by incorporating structure and position information. *Expert Systems with Applications*, *245*, 123140. https://doi.org/10.1016/j.eswa.2024.123140

APÊNDICE A

ONTOLOGIA PARA MINERAÇÃO TEXTUAL DE PATENTES EM LÍNGUA PORTUGUESA




OSF HOME
My Projects
Search
Support
Donate
Katia Cinara Tregnago Cunha
Funding/Support Information
Funder: Coordenação de Aperfeiçoamento de Pessoal de Nível Superior
Award title:
Award info URI:
Award number:
Funder: Conselho Nacional de Desenvolvimento Científico e Tecnológico
Award title:
Award info URI:
Award number:
Funder: Universidade Nove de Julho
Award title:
Award info URI:
Award number:
Affiliated institutions
Date created
April 20, 2024
Date modified
September 27, 2025
doi
Help

APÊNDICE B

SETUP EXPERIMENTAL

LEIA-ME Licença

License MIT

Projeto de Pesquisa de IA de Patentes

Este repositório contém o desenvolvimento de um projeto de pesquisa em Inteligência Artificial focado na análise e modelagem de dados relacionados a patentes. O objetivo é criar ferramentas, métodos e modelos que apoiem a extração, classificação e interpretação de informações tecnológicas de bancos de dados de patentes.

Estrutura do Projeto

Abaixo está a estrutura de diretórios proposta para organizar os artefatos de pesquisa de forma clara e produtiva:

```
data/
├── raw/ # Raw (unprocessed) data
├── processed/ # Cleaned and processed data
└── README.md # Description of the datasets

docs/ # Project documentation
├── README.md # Reading recommendations, drafts, etc.
├── meeting-notes/ # Meeting minutes

notebooks/ # Jupyter notebooks for analysis/experiments
├── exploratory/ # Exploratory analyses
├── experiments/ # Formal experiments
└── sandbox/ # Drafts and tests

src/ # Project source code
├── data/ # Data processing scripts
├── features/ # Feature engineering
├── models/ # Model development
└── utils/ # Common utilities

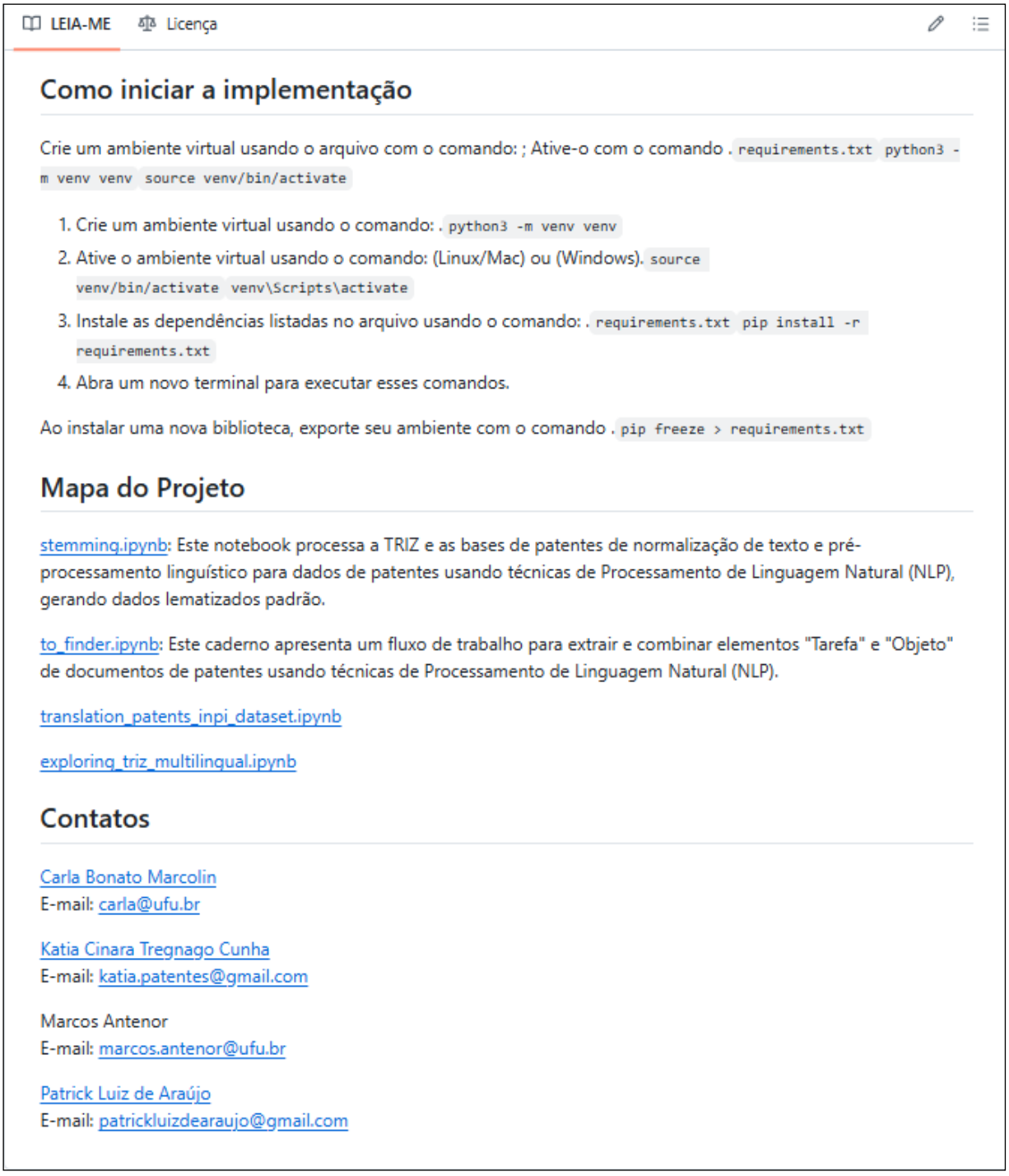
models/ # Trained and serialized models
└── README.md # Model descriptions

.gitignore # Files to be ignored by Git

LICENSE # Project license

README.md # Project overview (this file)

requirements.txt # Python dependencies
```



LEIA-ME Licença

Como iniciar a implementação

Crie um ambiente virtual usando o arquivo com o comando: ; Ative-o com o comando . `requirements.txt` `python3 -m venv venv` `source venv/bin/activate`

1. Crie um ambiente virtual usando o comando: . `python3 -m venv venv`
2. Ative o ambiente virtual usando o comando: (Linux/Mac) ou (Windows). `source venv/bin/activate` `venv\Scripts\activate`
3. Instale as dependências listadas no arquivo usando o comando: . `requirements.txt` `pip install -r requirements.txt`
4. Abra um novo terminal para executar esses comandos.

Ao instalar uma nova biblioteca, exporte seu ambiente com o comando . `pip freeze > requirements.txt`

Mapa do Projeto

stemming.ipynb: Este notebook processa a TRIZ e as bases de patentes de normalização de texto e pré-processamento linguístico para dados de patentes usando técnicas de Processamento de Linguagem Natural (NLP), gerando dados lematizados padrão.

to_finder.ipynb: Este caderno apresenta um fluxo de trabalho para extrair e combinar elementos "Tarefa" e "Objeto" de documentos de patentes usando técnicas de Processamento de Linguagem Natural (NLP).

translation_patents_inpi_dataset.ipynb

exploring_triz_multilingual.ipynb

Contatos

Carla Bonato Marcolin
E-mail: carla@ufu.br

Katia Cinara Tregnago Cunha
E-mail: katia.patentes@gmail.com

Marcos Antenor
E-mail: marcos.antenor@ufu.br

Patrick Luiz de Araújo
E-mail: patrickluizdearaujo@gmail.com

APÊNDICE C

LISTA DE PUBLICAÇÕES DA REVISÃO SISTEMÁTICA DA LITERATURA DO ESTUDO 1

1	General-use unsupervised keyword extraction model for keyword analysis
2	Patent Specialization for Deep Learning Information Retrieval Algorithms
3	Sustainable Product Innovation Using Patent Mining and TRIZ
4	The Software for Identifying Technological Complementarity Between Enterprises Based on Patent Databases
5	Intelligent technology management based on patent topic modeling
6	Natural Language Processing in assistance to Inventive Design activities
7	Patent Classification with Intelligent Keyword Extraction
8	Research on Patent Information Extraction Based on Deep Learning
9	DeepKEA: Employing Deep Learning Models for Keyword Extraction from Patent Documents
10	PAI-NET: Retrieval-Augmented Generation Patent Network Using Prior Art Information
11	SEA-PS: Semantic embedding with attention to measuring patent similarity by leveraging various text fields
12	Unsupervised technical phrase extraction by incorporating structure and position information
13	LLM-Based Extraction of Contradictions from Patents
14	A patent retrieval method and system based on double classification
15	A New Entity Relationship Extraction Method for Semi-Structured Patent Documents
16	Leveraging Information Retrieval Pipelines for Inventive Design: Application in Efficient Lattice Structures Manufacturing
17	Technological topic analysis of standard-essential patents based on the improved Latent Dirichlet Allocation (LDA) model
18	Development of a technology tree using patent information
19	SAO2Vec: Development of an algorithm for embedding the subject-action-object (SAO) structure using Doc2Vec
20	Patent Analysis Using an Ontology of Qualities of Inorganic Materials Based on Context-Dependency
21	An entity and relation extraction model based on context query and axial attention towards patent texts
22	Innovative product design based on radical problem solving
23	Patent Recommendation Methods Based on Transformer Encoders and Regularization Strategies
24	Application of machine learning in technological forecasting
25	Towards the extraction of semantic relations in design with natural language processing
26	Entity Identification of Patent Information for Carbon Capture Technology Based on the RoBERTa-BiLSTM-CRF Model
27	Technology opportunity discovery based on patent analysis: a hybrid approach of subject-action-object and generative topographic mapping
28	Comparing Complex Concepts with Transformers

29	Technology opportunity analysis using hierarchical semantic networks and dual link prediction
30	Patent-Based Technology Efficacy Information Extraction in Product Innovation Design
31	Enhancing semantic text similarity with functional semantic knowledge (FOP) in patents
32	From technology opportunities to solutions generation via patent analysis: Application of machine learning-based link prediction
33	A Novel Patent Knowledge Extraction Method for Innovative Design
34	Patent technical function-effect representation and mining method
35	Exploitation of Causal Relation for Automatic Extraction of Contradiction from a Domain-Restricted Patent Corpus
36	A Semantic Understanding Method for Patent Text Based on Large Language Model
37	Research on Product Core Component Acquisition Based on Patent Semantic Network
38	Utilization of bio-ontologies for enhancing patent information retrieval
39	Intelligent compilation of patent summaries using machine learning and natural language processing techniques
40	Patent landscape and key technology interaction roadmap using graph convolutional network - Case of mobile communication technologies beyond 5G
41	TechWordNet: Development of semantic relation for technology information analysis using F-term and natural language processing
42	Identifying technology opportunity using SAO semantic mining and outlier detection method: A case of triboelectric nanogenerator technology
43	TechNet: Technology semantic network based on patent data
44	Semantic Based Heterogeneous Information Network Embedding for Patent Citation Recommendation
45	Combining topic modeling and SAO semantic analysis to identify technological opportunities of emerging technologies
46	Systematic Review on Identification and Prediction of Deep Learning Based Cyber Security Technology and Convergence Fields
47	Intelligent Text Mining for Ontological Knowledge Graph Refinement and Patent Portfolio Analysis—Case Study of Net-Zero Data Center Innovation Management
48	Exploring the Evolution of Core Technologies in Agricultural Machinery: A Patent-Based Semantic Mining Analysis
49	Automatic extraction of inventive information out of patent texts in support of manufacturing design studies using Natural Languages Processing
50	Semi-automatic extraction of technological causality from patents
51	Methods for extracting the descriptions of sci-tech effects and morphological features of technical systems from patents
52	A technical patent map construction method and system based on multi-dimensional technical feature extraction
53	Ontology-based patent analysis for bike-sharing services: Identifying competitive advantages of a product-service business
54	Integrating Technology-Relationship-Technology Semantic Analysis and Technology Roadmapping Method: A Case of Elderly Smart Wear Technology
55	Automatic Construction of Technology Function Matrix
56	Big Data Related Patent Retrieval System Based on Filtering Rules
57	A patent keywords extraction method using TextRank model with prior public

	knowledge
58	Optimizing Patent Prior Art Search: An Approach Using Patent Abstract and Key Terms
59	Building knowledge graphs from technical documents using named entity recognition and edge weight updating neural network with triplet loss for entity normalization
60	Concept Extraction Based on Semantic Models Using Big Amount of Patents and Scientific Publications Data
61	Discovery of topic evolution path and semantic relationship based on patent entity representation
62	Patent keyword extraction algorithm based on distributed representation for patent classification
63	TechPat: Technical Phrase Extraction for Patent Mining
64	Technology identification from patent texts: A novel named entity recognition method
65	Patent Keyword Analysis Using Regression Modeling Based on Quantile Cumulative Distribution Function
66	Open Relation Extraction in Patent Claims with a Hybrid Network
67	AI for Patents: A Novel Yet Effective and Efficient Framework for Patent Analysis
68	A simple and fast method for Named Entity context extraction from patents
69	Question Answering with Transformers and Few-Shot Learning to Find Inventive Solutions for IDM-TRIZ Problems and Contradictions in Patents
70	Unveiling the inventive process from patents by extracting problems, solutions and advantages with natural language processing
71	Unveiling Technological Evolution with a Patent-Based Dynamic Topic Modeling Framework: A Case Study of Advanced 6G Technologies
72	Discovering new business opportunities with dependent semantic parsers
73	Comparing text corpora via topic modelling
74	Patent Text Classification based on Deep Learning and Vocabulary Network
75	Ontology-based heuristic patent search
76	Intelligent extraction of a knowledge ontology from global patents: The case of smart retailing technology mining
77	Analyzing technological competencies in the patent-based supplier portfolio: introducing an approach for supplier evaluation using semantic anchor points and similarity measurements
78	Technological trend mining: identifying new technology opportunities using patent semantic analysis
79	Exploring Technology Opportunities Based on User Needs: Application of Opinion Mining and SAO Analysis
80	Automatic users extraction from patents
81	Detecting Multi Word Terms in patents the same way as entities
82	Study on the Technology Trend Screening Framework Using Unsupervised Learning
83	Automating the search for a patent's prior art with a full text similarity search
84	A deep learning based method benefiting from characteristics of patents for semantic relation classification
85	DeepPatent: patent classification with convolutional neural networks and word embedding
86	A patent text classification model based on multivariate neural network fusion

87	Query generation for patent retrieval with keyword extraction based on syntactic features
88	Technical Phrase Extraction for Patent Mining: A Multi-level Approach
89	Investigating technology opportunities: the use of SAOx analysis
90	A deep learning based method for extracting semantic information from patent documents
91	Early detection of technology opportunity based on analogy design and phrase semantic representation
92	Exploiting word embedding for heterogeneous topic model towards patent recommendation
93	Doc2vec-based link prediction approach using SAO structures: application to patent network
94	Chinese technical terminology extraction based on DC-value and information entropy
95	Improved Technology Similarity Measurement in the Medical Field based on Subject-Action-Object Semantic Structure: A Case Study of Alzheimer's Disease
96	Using supervised machine learning for large-scale classification in management research: The case for identifying artificial intelligence patents
97	Technical problem identification for supervised state of the art
98	An integrated implicit user preference mining approach for uncertain conceptual design decision-making: A pipeline inspection trolley design case study
99	Method of identification of patent trends based on descriptions of technical functions
100	PaEffExtr: A Method to Extract Effect Statements Automatically from Patents
101	Research on Patent Text Classification Based on Word2Vec and LSTM
102	Similar patent search method based on a functional information fusion
103	Text Simplification of Patent Documents
104	The method for detecting the dependencies between technical functions and physical effects
105	Using non-supervised machine learning to generate a knowledge ontology for patent analytics
106	Construction of a Matrix "Physical Effects - Technical Functions" on the Base of Patent Corpus Analysis
107	Engineering knowledge graph for keyword discovery in patent search
108	Identification of descriptions of scientific-technical effects in patent documents
109	Named Entity Recognition for Biomedical Patent Text using Bi-LSTM Variants
110	Patent Quality Classification System Using the Feature Extractor of Deep Recurrent Neural Network
111	Relation Extraction Toward Patent Domain Based on Keyword Strategy and Attention+BiLSTM Model
112	Improvement of Automatic Extraction of Inventive Information with Patent Claims Structure Recognition
113	TRIZ Technical Contradiction Extraction Method Based on Patent Semantic Space Mapping
114	A Patent recommendation algorithm based on topic classification and semantic similarity
115	Automatic Construction of a Domain-specific Knowledge Graph for Chinese Patent Based on Information Extraction
116	Automatic Extraction of Potentially Contradictory Parameters from Specific Field Patent Texts

117	Linguistically informed masking for representation learning in the patent domain

APÊNDICE D

LISTA DE PUBLICAÇÕES DA REVISÃO SISTEMÁTICA DA LITERATURA DO ESTUDO 2

1	PaTRIZ: A framework for mining TRIZ contradictions in patents
2	A text-mining-based patent network: Analytical tool for high-technology trend
3	Patent analysis for systematic innovation: Automatic function interpretation and automatic classification of level of invention using natural language processing and artificial neural networks
4	An exploration study of construction innovation principles: Comparative analysis of construction scaffold and template patents
5	AI Based Patent Analyzer for Suggesting Solutive Actions and Graphical Triggers During Problem Solving
6	Toward an automatic extraction of idm concepts from patents
7	Using text mining to retrieve information about circular economy
8	Identifying technology trends for R and D planning using TRIZ and text mining
9	Development of innovative product design process using patent multi-scale analysis and TRIZ methodology,
10	Automatic Extraction of Potentially Contradictory Parameters from Specific Field Patent Texts
11	Identification of promising patents for technology transfers using TRIZ evolution trends
12	Extraction of physical effects based on the semantic analysis of the patent texts
13	Patent Specialization for Deep Learning Information Retrieval Algorithms
14	PatRIS: Patent Ranking Inventive Solutions
15	Computer-aided comparison of thesauri extracted from complementary patent classes as a means to identify relevant field parameters
16	Searching for similar products through patent analysis
17	Method of detection of technical functions performed by physical effects
18	Knowledge based approach for formulating TRIZ contradictions
19	SAO extraction on patent discovery system development for Islamic Finance and Banking
20	Extraction and linking of motivation, specification and structure of inventions for early design use
21	A new function-based patent knowledge retrieval tool for conceptual design of innovative products
22	Integrated model for technology assessment and expected evolution: A case study in the chilean mining industry
23	A fact-oriented ontological approach to SAO-based function modeling of patents for implementing Function-based Technology Database
24	Setting Up Context-Sensitive Real-Time Contradiction Matrix of a Given Field Using Unstructured Texts of Patent Contents and Natural Language Processing
25	Conceptual Semantic Analysis of Patents and Scientific Publications Based on TRIZ Tools
26	Construction of a Matrix “Physical Effects – Technical Functions” on the Base of

	Patent Corpus Analysis
27	The contribution of syntactic-semantic approach to the search for complementary literatures for scientific or technical discovery
28	Automatic extraction of inventive information out of patent texts in support of manufacturing design studies using Natural Languages Processing,
29	SummaTRIZ : Summarization Networks for Mining Patent Contradiction
30	Starting from patents to find inputs to the Problem Graph model of IDM-TRIZ
31	ARIZ85 and patent-driven knowledge support
32	Discovering shifts in competitive strategies in probiotics, accelerated with TechMining
33	Question Answering with Transformers and Few-Shot Learning to Find Inventive Solutions for IDM-TRIZ Problems and Contradictions in Patents
34	SynCRF: Syntax-Based Conditional Random Field for TRIZ Parameter Minings
35	A smart conflict resolution model using multi-layer knowledge graph for conceptual design
36	The development of a modified TRIZ Technical System ontology
37	Three-steps methodology for patents prior-art retrieval and structured physical knowledge extracting
38	Patent analysis with text mining for TRIZ
39	On classification and extraction of deep knowledge in patents based on TRIZ Theory
40	Extraction of Principle Knowledge from Process Patents for Manufacturing Process Innovation
41	Agile innovation process model based on computer-aided patent knowledge mining and functional analogy
42	Semi-automatic extraction of technological causality from patents
43	A new TRIZ-based patent knowledge management system for construction technology innovation
44	Summarization as a Denoising Extraction Tool
45	Extraction of physical effects practical applications from patent database
46	Design of the patent evasion and rescue manipulator based on TRIZ
47	Natural Language Processing (NLP) - A solution for knowledge extraction from patent unstructured data
48	Using patents to populate an inventive design ontology
49	Exploitation of Causal Relation for Automatic Extraction of Contradiction from a Domain-Restricted Patent Corpus
50	Combining tech-mining and semantic-TRIZ for a faster and better technology analysis: A case in energy storage systems
51	The method for detecting the dependencies between technical functions and physical effects
52	Computer-aided classification of patents oriented to TRIZ
53	Identification of Product Innovation Path Incorporating the FOS and BERTopic Model from the Perspective of Invalid Patents
54	Improvement of Automatic Extraction of Inventive Information with Patent Claims Structure Recognition
55	Text Simplification of Patent Documents
56	Functional-based search for patent technology transfer

57	A new method for extracting knowledge from patents to inspire designers during the problem-solving phase
58	Innovation Logic: Benefits of a TRIZ-Like Mind in AI Using Text Analysis of Patent Literature
59	A Conceptual Design Framework based on TRIZ Scientific Effects and Patent Mining
60	Key Technologies for Sustainable Design Based on Patent Knowledge Mining
61	Replicating TRIZ Reasoning Through Deep Learning
62	A function-behaviour oriented search for patent digging
63	Concept Extraction Based on Semantic Models Using Big Amount of Patents and Scientific Publications Data.
64	A lexico-syntactic pattern matching method to extract IDM- TRIZ knowledge from on-line patent databases
65	Construction and Application of Enterprise Knowledge Base for Product Innovation Design
66	Interrelating products through properties via patent analysis
67	Computer-aided analysis of patents and search for TRIZ contradictions
68	Classification of TRIZ Inventive Principles and Sub-principles for Process Engineering Problems
69	An SAO-based approach to patent evaluation using TRIZ evolution trends
70	TrendPerceptor: A property-function based technology intelligence system for identifying technology trends from patents
71	TRIZ Technical Contradiction Extraction Method Based on Patent Semantic Space Mapping
72	Unveiling Technological Evolution with a Patent-Based Dynamic Topic Modeling Framework: A Case Study of Advanced 6G Technologies
73	LLM-Based Extraction of Contradictions from Patents
74	AI-Aided Resource Mining Method for Idealization-Driven Product Innovation
75	QFD and TRIZ integration in product development: A Model for Systematic Optimization of Engineering Requirements

APÊNDICE E

CARTA-CONVITE AOS ESPECIALISTAS DO ESTUDO 4

Assunto: Convite para participação em pesquisa acadêmica

Prezado(a) [Nome do Especialista],

Meu nome é Kátia Cinara Tregnago Cunha, doutoranda em Administração na Universidade Nove de Julho. Estou conduzindo a pesquisa intitulada *"Mineração Textual de Inteligência Técnica de Patentes em Língua Portuguesa a partir de uma Ontologia Baseada na Teoria da Resolução Inventiva de Problemas (TRIZ)"*, cujo objetivo é desenvolver um método de mineração textual de inteligência técnica em documentos de patente redigidos em língua portuguesa, utilizando uma ontologia fundamentada na TRIZ em conjunto com métodos analíticos de Inteligência Artificial.

O estudo é orientado pela Profa. Dra. Cristina Dai Prá Martens (Universidade Nove de Julho), com coorientação da Profa. Dra. Carla Bonato Marcolin (Universidade Federal de Uberlândia) e do Prof. Dr. Carlos J. Costa (Instituto Superior de Economia e Gestão da Universidade de Lisboa).

Considerando sua experiência e reconhecida atuação nas áreas de Engenharia/Farmácia e Propriedade Intelectual, acreditamos que sua participação será de grande relevância para enriquecer e aprofundar os resultados da pesquisa.

O instrumento de coleta consistirá em um formulário contendo cerca de 50 títulos e resumos de patentes obtidos junto ao Instituto Nacional da Propriedade Industrial (INPI). Esses documentos foram previamente processados com ferramentas de Inteligência Artificial, a partir da ontologia TRIZ por mim desenvolvida, a fim de extrair termos representativos capazes de identificar a função, o objeto e o efeito físico aplicado, sugerindo potenciais soluções genéricas para diferentes problemas técnicos.

Sua participação consistirá em avaliar se os termos extraídos refletem adequadamente o conteúdo descrito nos títulos e resumos das patentes e se estão alinhados aos efeitos físicos e às respectivas categorias de tarefa e objeto definidos na ontologia TRIZ. A avaliação será realizada por meio de perguntas fechadas, com duas opções de resposta. O tempo estimado para o preenchimento é de até 5 horas, e o prazo para devolutiva é de até 30 dias.

O formulário conterá instruções detalhadas para o preenchimento, bem como acesso à ontologia para consulta. Tanto as respostas quanto a identidade dos participantes serão mantidas em anonimato.

A participação é voluntária e poderá ser interrompida a qualquer momento. Caso tenha interesse em participar, solicito a gentileza de confirmar seu retorno para que possamos agendar uma reunião remota, com duração aproximada de 30 minutos, na qual apresentarei a dinâmica de preenchimento do formulário e a ontologia TRIZ.

Agradeço antecipadamente sua atenção e colaboração.

Atenciosamente,

Kátia Cinara Tregnago Cunha
Programa de Pós-graduação em Administração – Universidade Nove de Julho
✉ katia.patentes@gmail.com
(51) 99172-2817

APÊNDICE F
FORMULÁRIO DE AVALIAÇÃO DO ESTUDO 4

Seção 1 de 3

AVALIAÇÃO DE RECUPERAÇÃO DE INTELIGÊNCIA TÉCNICA DE PATENTES

B *I* U

Este formulário foi desenvolvido para avaliar a qualidade e consistência dos termos atribuídos a cada patente através de um método de mineração textual baseado em Inteligência Artificial (IA) e que utiliza uma ontologia como base semântica.

A ontologia desenvolvida sistematiza conceitos tecnológicos e suas relações, incorporando padrões linguísticos (verbo/substantivo) que orientam a ferramenta de IA na aquisição de padrões e geração de dados. Esta ontologia reúne efeitos físicos derivados da Teoria da Resolução Inventiva de Problemas (*TRIZ*), contemplando um relacionamento ternário entre as subclasses Tarefa (T), Objeto (O) e Efeitos Físicos. A Ontologia está disponível em repositório de acesso aberto.

As patentes analisadas foram obtidas da base do INPI/Brasil, sendo disponibilizados os campos textuais de título e resumo.

Principais conceitos:

- Tarefa (T) - Ação potencial expressa por verbo. A tarefa deve refletir o problema que a patente resolve.
- Objeto (O) - Meio pelo qual uma tarefa atinge seu propósito, expresso por substantivo (categorizado como: sólido, sólido dividido, líquido ou gás). O objeto pode ser um dispositivo, substância, método, sistema, entre outros.
- Efeito Físico (EF) - Resultado da aplicação de uma Tarefa em um Objeto, representando soluções genéricas validadas cientificamente. Procure por fenômenos físicos, químicos ou biológicos.
- Termos atribuídos - relacionamento ternário (T, O, EF) atribuído pela ferramenta de IA, que pode ser obtido da Ontologia ou extraído do texto analisado quando não encontrado um conjunto de termos similares.
- Termos derivados - termos originais da Ontologia. Quando não presentes, significa que os termos atribuídos pela ferramenta de IA são originais.

O que esperamos de você:

1. Verificar se os relacionamentos Tarefa e Objeto atribuídos a cada patente são compatíveis com o conteúdo do título e com o conteúdo do resumo. Define-se compatibilidade a proximidade semântica dos termos T ou O com o conteúdo textual analisado.

2. Verificar se o relacionamento ternário (Tarefa–Objeto–Efeito Físico), definido na Ontologia ou dela derivado e atribuído a cada patente, fornece uma sugestão de solução técnica que possa ser respondida pela própria patente.

Informações da Patente BR102020003953-9

- **Título:** COMPOSIÇÃO ADESIVA DE CONTATO BASE ÁGUA
- **Resumo:**
  É descrita uma composição adesiva de contato base água que compreende uma dispersão de etileno e acetato de vinila (VAE) associada com resinas taquificantes e um nanoparticulado mineral, além do uso de agente coalescente e cosolvente para a dispersão das resinas, evitando o uso de plastificante à base ftalatos ou outros solventes perigosos, com a finalidade de alcançar as mesmas propriedades das composições adesivas à base de solvente.
- **Termos da Ontologia atribuídos:**
  - **Tarefa:** Adesivar
  - **Objeto:** Superfície
  - **Efeito físico:** Aderência

Termos derivados:
Tarefa: Aquecer
Objeto: Líquido
Efeito físico: Adesivo

Q1 - As subclasses Tarefa (T) e Objeto (O) atribuídas à patente são compatíveis com o **título** da patente? *

	SIM [T e/ou O aparecem explicitamente ou de forma inferível no campo textual analisado.	NÃO [T e/ou O não têm relação e não são inferíveis no campo textual analisado.	Indeterminado [as informações do campo textual analisado não são suficientes.	Indeterminado [o relacionamento T–O é demasiado genérico.]
Compatibilidade com o título	○	○	○	○
Compatibilidade com o resumo	○	○	○	○

Q2 - As subclasses Tarefa (T) e Objeto (O) atribuídas à patente são compatíveis com o **resumo** da patente? *

	SIM [T e/ou O aparecem explicitamente ou de forma inferível no campo textual analisado.	NÃO [T e/ou O não têm relação e não são inferíveis no campo textual analisado.	Indeterminado [as informações do campo textual analisado não são suficientes.	Indeterminado [o relacionamento T–O é demasiado genérico.]
Compatibilidade com o título	○	○	○	○
Compatibilidade com o resumo	○	○	○	○

Q3 - O relacionamento ternário atribuído à patente fornece uma sugestão de solução técnica que é efetivamente abordada ou resolvida pela patente? *

- [ ] SIM
- [ ] NÃO

Q4 - Existe consistência entre os termos derivados (T, O, EF) e aqueles dos quais se originam? (responda somente se a coluna "derivado do relacionamento semântico da Ontologia" estiver preenchida).

- [ ] SIM
- [ ] NÃO