REVISTA BRASILEIRA DE DESENVOLVIMENTO E INOVAÇÃO

ANÁLISE DE SIMILARIDADE SEMÂNTICA DE PATENTES UTILIZANDO PROCESSAMENTO DE LINGUAGEM NATURAL E BUSCA HÍBRIDA

ISABELA AGUIAR[1]; JOSÉ LAURINDO CAMPOS DOS SANTOS[2]; ANDRÉA CORRÊA FLÔRES ALBUQUERQUE[3]

RESUMO: Este estudo apresenta o desenvolvimento e a validação de um sistema automatizado para triagem preliminar de patentes, fundamentado na análise semântica do campo de reivindicações (*claims*), seção que define o escopo técnico-jurídico da proteção. O sistema foi implementado em *Python* 3.11, utilizando técnicas de Processamento de Linguagem Natural (PLN), que permitem a interpretação computacional de textos, e *embeddings*, representações numéricas que possibilitam mensurar similaridade conceitual entre documentos. O modelo adotado, paraphrase-multilingual-mpnet-base-v2, foi integrado a uma base local e à API da Lens.org, que reúne milhões de patentes de múltiplas jurisdições. Na interface desenvolvida em *Streamlit*, o usuário insere o texto das *claims* e recebe uma lista ranqueada com as dez patentes mais semelhantes, acompanhadas de título, link e percentual de similaridade. A avaliação foi conduzida com 100 pares de patentes rotulados manualmente. Os resultados indicaram alto desempenho em situações claras e robustez consistente em cenários ambíguos. Conclui-se que a solução contribui para a eficiência, a padronização e a confiabilidade da triagem, reduzindo a sobrecarga de especialistas e ampliando a assertividade na proteção de ativos intangíveis em Núcleos de Inovação Tecnológica e departamentos de Propriedade Intelectual.

PALAVRAS-CHAVE: Propriedade Intelectual. Reivindicações de Patentes. Busca Semântica. Processamento Multilíngue. *Embeddings*.

SEMANTIC SIMILARITY ANALYSIS OF PATENTS USING NATURAL LANGUAGE PROCESSING AND HYBRID SEARCH

ABSTRACT: *This study presents the development and validation of an automated system for preliminary patent screening, based on the semantic analysis of the claims section, which defines the technical-legal scope of protection. The system was implemented in Python 3.11, using Natural Language Processing (NLP) techniques that enable the computational interpretation of texts, and embeddings, numerical representations that make it possible to*

[1] Estudante de Engenharia da Computação, Centro Universitário FAMETRO, Manaus, Amazonas, Brasil. ORCID: 0009-0002-6305-8085. E-mail: isabela.andradeaguiar@gmail.com.

[2] Ph.D. em Ciência da Computação. Instituto Nacional de Pesquisas da Amazônia. Manaus, Amazonas, Brasil. ORCID: 0000-0001-5363-8225. E-mail: laurindo.campos@inpa.gov.br.

[3] Dra. em Informática. Instituto Nacional de Pesquisas da Amazônia. Manaus, Amazonas, Brasil. ORCID: 0000-0003-3513-6074. E-mail: andreaalb.1993@gmail.com.

measure conceptual similarity between documents. The chosen model, paraphrase-multilingual-mpnet-base-v2, was integrated with both a local database and the Lens.org API, which aggregates millions of patents from multiple jurisdictions. In the Streamlit-based interface, the user inputs the claims text and receives a ranked list of the ten most similar patents, along with title, link, and similarity percentage. The evaluation was conducted with 100 manually labeled patent pairs. The results indicated high performance in clear-cut cases and consistent robustness in ambiguous scenarios. It is concluded that the solution contributes to the efficiency, standardization, and reliability of screening, reducing the workload of specialists and increasing accuracy in the protection of intangible assets within Technology Innovation Centers and Intellectual Property departments.

Key words: *Intellectual Property; Claims; Semantic Search; Multilingual Processing; Embeddings.*

INTRODUÇÃO

A propriedade intelectual (PI) constitui um instrumento estratégico para o estímulo à inovação tecnológica e à valorização de ativos intangíveis. Por meio de dispositivos legais como patentes, marcas e segredos industriais, os sistemas de PI promovem a proteção de invenções, a difusão do conhecimento técnico e a competitividade econômica em escala global (WIPO, 2022).

No Brasil, a proteção dos direitos de PI é regulamentada pela Lei nº 9.279/1996, sendo o Instituto Nacional da Propriedade Industrial (INPI) responsável por sua execução. Em instituições públicas de pesquisa, como universidades e institutos tecnológicos, a gestão da PI envolve desafios operacionais significativos, especialmente no que se refere à triagem e à submissão de pedidos de patente. A avaliação da viabilidade técnica de uma invenção exige a análise criteriosa do campo de reivindicações, a análise comparativa com o conhecimento técnico já disponível e a interpretação jurídica, etapas que demandam alta especialização e tempo considerável (WIPO, 2023). Esse cenário compromete a eficiência institucional, sobretudo diante da escassez de recursos humanos qualificados para examinar grandes volumes de pedidos.

As reivindicações representam a seção central de um pedido de patente, pois delimitam, de forma técnica e legal, o escopo exato da proteção requerida. Diferentemente da descrição ou do resumo, que apenas contextualizam a invenção, as reivindicações definem juridicamente o que se pretende proteger com exclusividade. O uso de campos textuais completos, como as reivindicações, é reconhecido como essencial para análises de similaridade técnica, conforme

discutido por Tseng et al. (2007), que destacam o papel da mineração de texto na recuperação de patentes tecnicamente correlatas. Por essa razão, constituem o principal foco tanto de processos de concessão quanto de litígios. Assim, qualquer iniciativa de automação na análise de patentes deve priorizar esse campo como elemento central de comparação técnica.

Nos últimos anos, o avanço das técnicas de inteligência artificial, especialmente no campo do processamento de linguagem natural (PLN), tem possibilitado a aplicação de modelos capazes de capturar o significado semântico de textos técnicos com alta precisão. Dentre essas técnicas, destaca-se o uso de *embeddings*, representações vetoriais densas que convertem o conteúdo textual em vetores numéricos, permitindo a comparação entre documentos com base em sua similaridade semântica, mesmo quando redigidos com terminologias distintas (Reimers; Gurevych, 2019). Essa abordagem mostra-se particularmente útil no contexto de patentes, no qual a diversidade na redação técnica pode ocultar proximidades conceituais relevantes.

Para viabilizar esse tipo de análise, é essencial o acesso a bases de dados técnicas estruturadas e abrangentes. A plataforma *The Lens.org* oferece uma das infraestruturas mais completas e acessíveis, integrando milhões de documentos de patentes de múltiplas jurisdições (USPTO, EPO, WIPO, CIPO) e disponibilizando o conteúdo integral do campo reivindicações com suporte a APIs (Penfold, 2020). Além dela, destacam-se o *Espacenet*, mantido pelo Escritório Europeu de Patentes (EPO), com mais de 140 milhões de documentos; o *Google Patents*, que fornece ferramentas de busca semântica e tradução automática; a *PatentsView*, voltada à análise estatística de patentes dos EUA; e o WIPO *Patentscope*, que agrega pedidos internacionais sob o Tratado de Cooperação em Matéria de Patentes (PCT), com recursos multilíngues. Essas plataformas fornecem a base necessária para a aplicação de técnicas modernas de IA em tarefas como triagem, recomendação e avaliação técnica automatizada.

Neste contexto, este trabalho apresenta um sistema automatizado de triagem preliminar de patentes, desenvolvido com o objetivo de atuar como filtro técnico inicial no processo de submissão institucional. A solução emprega modelos multilíngues de *embeddings* para calcular a similaridade semântica entre o campo de reivindicações de um pedido e uma base técnica previamente indexada, retornando os documentos mais similares. O sistema foi validado com base em um conjunto rotulado manualmente, apresentando desempenho robusto em métricas como acurácia, *F1-score*, MCC e AUC. Ao oferecer uma triagem automatizada, padronizada e interpretável, essa abordagem representa uma contribuição técnica relevante no apoio à decisão em propriedade intelectual, promovendo maior eficiência, agilidade e consistência na gestão

institucional de inovações.

CONTEXTO INSTITUCIONAL E LACUNAS

O INPA opera em um ambiente de alta produção de P&D, no qual resultados que vão da pesquisa aplicada ao desenvolvimento experimental precisam ser rapidamente triados para decidir o que depositar, quando e como proteger. Na prática, os núcleos de PI convivem com *backlog* e carga analítica elevada, o que torna a triagem técnico-documental lenta e heterogênea entre áreas. Essa variabilidade afeta a qualidade dos pareceres internos, alonga ciclos decisórios e pode comprometer prazos estratégicos (BRASIL, 1996; OMC, 1994).

A lacuna central é a ausência de um procedimento de triagem padronizado, reprodutível e rápido para: (i) verificar similaridade técnica (*prior art*), (ii) qualificar o quadro de reivindicações e (iii) organizar evidências para respostas a exigências.

Objetivo e contribuições. O artigo apresenta e avalia uma abordagem de apoio à triagem técnica que acelera a busca de similaridade, uniformiza pareceres e prioriza casos por risco/impacto para decisões de portfólio.

MARCO LEGAL ESSENCIAL E FLUXO DE PATENTES NO BRASIL (FOCO OPERACIONAL)

No Brasil, as etapas e prazos do processamento de patentes seguem a Lei nº 9.279/1996 (LPI), em harmonia com padrões mínimos do TRIPS/ADPIC (BRASIL, 1996; OMC, 1994). Operacionalmente, o rito no INPI encadeia-se assim: o pedido é depositado (relatório descritivo, reivindicações, desenhos e resumo) e passa por exame formal; permanece sob sigilo por até 18 meses a partir da data de depósito ou da prioridade mais antiga e, então, é publicado (BRASIL, 1996, art. 30). O requerimento de exame deve ser feito em até 36 meses do depósito, sob pena de arquivamento (BRASIL, 1996, art. 33). Inicia-se a fase de exame técnico, com busca de anterioridade e análise dos requisitos de patenteabilidade; podem ocorrer exigências com manifestações do depositante. Ao final, o INPI profere decisão (deferimento/indeferimento), cabendo recurso administrativo nas hipóteses legais; em caso de concessão, expede-se a carta-patente e o titular mantém o direito mediante anuidades (BRASIL, 1996).

Onde a ferramenta ajuda no fluxo. Há três pontos de inserção prioritários: (a) pré-

depósito, para checagem de similaridade e sanidade do quadro de reivindicações; (b) pré-exame, para priorizar casos por risco/impacto (classe CPC, probabilidade de anterioridade, valor estratégico); e (c) resposta a exigências, para organizar referências e argumentos técnicos com rastreabilidade.

CRONOLOGIA DO PROBLEMA (2020–2025)

Entre 2020 e 2021, o INPI consolidou a execução do Plano de Combate ao *Backlog* iniciado em 2019, com acompanhamento público da redução do estoque. Em 2022, houve melhoria do *throughput* de exame em relatórios institucionais. Em 2023, a agenda incluiu automação de etapas do fluxo e ampliações de PPH para acelerar decisões. Em 2024–2025, as metas de redução de prazos se mantiveram no planejamento 2023–2026, com continuidade do monitoramento do *backlog*; em paralelo, a OMPI reporta retomada gradual dos depósitos no Brasil e manutenção de prazos médios ainda desafiadores, reforçando a necessidade de eficiência na triagem (INPI, 2023; INPI, 2019–2025; OMPI, 2025).

Leitura prática dos prazos. Os marcos de 18 meses (publicação) e 36 meses (requerimento de exame) delimitam janelas críticas para coleta de evidências e preparação de pareceres. Por isso, um procedimento de triagem ágil e padronizado agrega valor antes do depósito, entre a publicação e o exame e durante respostas a exigências (BRASIL, 1996).

TRABALHOS RELACIONADOS

A literatura recente consolida quatro linhas principais para busca e análise de patentes: (i) lexical (e.g., BM25/TF-IDF), (ii) densa (*embeddings*), (iii) híbrida (combinações lexical + densa) e (iv) *reranking* com *cross-encoders*. Dois levantamentos abrangentes mapeiam tarefas, conjuntos de dados e tendências, incluindo desafios específicos do domínio de patentes, como estrutura textual, comprimento, linguagem jurídica e multimodalidade, e servem como síntese atualizada do estado da arte (Jiang; Goetz, 2025; Shomee et al., 2025).

No eixo lexical, BM25 continua competitivo em bases extensas e é frequentemente empregado como mecanismo de recuperação inicial em pipelines de múltiplos estágios. Levantamentos sistemáticos recentes sobre busca de anterioridade reforçam o predomínio de coleções baseadas no USPTO e nos *tracks* CLEF-IP, demonstrando a centralidade do

componente lexical na geração da lista candidata (Ali et al., 2024).

No eixo denso, modelos *Sentence-BERT* adaptados ao domínio de patentes elevaram a precisão semântica. O modelo *PatentSBERTa*, por exemplo, é treinado diretamente sobre reivindicações, validando sua aplicação em tarefas de similaridade entre patentes e classificação multi-rótulo com ganhos substanciais (Bekamiri; Hain; Jurowetzki, 2024). Esses trabalhos mostram ainda que restrições de contexto e custo computacional exigem arquiteturas eficientes para documentos longos, condição comum em patentes (Jiang; Goetz, 2025).

As abordagens híbridas integram recuperação lexical e vetorial, seguidas por *reranking* com *cross-encoders* (por exemplo, BERT) para reordenamento de um top-k de candidatos. Quando múltiplos sinais de ranqueamento são combinados, métodos como o *Reciprocal Rank Fusion* (RRF) mostram-se simples e eficazes para fusão de listas (Nogueira; Cho, 2019; Cormack; Clarke; Buettcher, 2009). Esses arranjos têm sido amplamente utilizados em sistemas de recuperação de informação e se mostram particularmente promissores no domínio de patentes, dada a complexidade das consultas e a densidade técnica dos documentos (Ali et al., 2024).

Quanto ao foco textual, estudos de classificação em larga escala indicam que as reivindicações, isoladamente, são suficientes para discriminar categorias CPC/IPC. Essa evidência sustenta o uso exclusivo desse campo em sistemas de triagem e busca técnica (Lee; Hsiang, 2020). Modelos densos recentes, como o próprio *PatentSBERTa*, também priorizam o campo de reivindicações para a geração de *embeddings* mais informativos (Bekamiri; Hain; Jurowetzki, 2024).

LACUNAS IDENTIFICADAS

(a) Idioma e dados: as coleções padrões de avaliação (ex: CLEF-IP) concentram-se em idiomas como inglês, francês e alemão, não contemplando o português, o que representa uma limitação importante para aplicações em contextos brasileiros e amazônicos (Piroi et al., 2011; Piroi et al., 2013). Além disso, os levantamentos recentes revelam o uso predominante de dados oriundos do USPTO/CLEF-IP, com escassa evidência de corpora pt-BR (Ali et al., 2024).

(b) Reprodutibilidade e comparabilidade: diferenças nos conjuntos de dados e nos protocolos metodológicos dificultam comparações diretas entre estudos, um problema apontado tanto nos levantamentos quanto em trabalhos aplicados (Jiang; Goetz, 2025; Bekamiri; Hain;

Jurowetzki, 2024).

MÉTODOS

Estudo aplicado, exploratório e quantitativo para desenvolver, implementar e validar um sistema automatizado de triagem técnica de patentes baseado na análise semântica do campo de *claims*. Adotamos boas práticas de reprodutibilidade: documentação de ambiente, fixação de versões, scripts executáveis, logs e congelamento de snapshot de dados para avaliação.

AMBIENTE COMPUTACIONAL

a) Linguagem: Python 3.11.

b) Principais bibliotecas: *Sentence-Transformers* (inferência do modelo), *torch*, *numpy*, *pandas*, nltk (limpeza/stopwords), *scikit-learn* (métricas/gráficos), *matplotlib* e *seaborn* (visualização), *streamlit* (*UI*).

c) Controle de versões e reprodutibilidade: versões fixadas em *requirements.txt*; os *scripts* de execução estão versionados (ex.: *make* evaluate, python -m *scripts*/avaliar.py), e a avaliação usa um *snapshot* de dados congelado (vide § CARTÃO DO CONJUNTO DE DADOS LOCAL). Como apenas realizamos inferência (sem treino), *seeds* não afetam os *embeddings*; ainda assim, os *random states* dos experimentos analíticos são fixados nos *scripts* de avaliação.

FONTES E COLETA DE DADOS (LENS.ORG)

a) Fonte principal: *API* pública da *Lens.org*, com cobertura multi-jurisdição (USPTO, EPO, WIPO, INPI, entre outras) e campo *claims* estruturado.

b) Arquitetura de ingestão: coleta incremental, paginação nativa da *API* e de duplicação por identificador.

c) Filtros efetivamente aplicados: inclusão apenas de registros com campo *claims* não vazio; sem filtro temporal ou de idioma na coleta (visando diversidade).

d) Persistência: armazenamento local em *JSON*/*Parquet* com metadados de fonte, jurisdição e carimbo de *data*/hora da coleta.

e) *Search* híbrida: (i) consultas na base local indexada; (ii) consultas *on-the-fly* à *API* da *Lens.org* com fusão por similaridade.

f) Reprodutibilidade: o script de coleta encontra-se no repositório (ex.: coleta_patentes.py), com parâmetros e exemplos de uso. *logs* de execução e *hash*es do lote coletado são mantidos.

CARTÃO DO CONJUNTO DE DADOS LOCAL (*DATASET CARD*)

a) Tamanho e composição: base local com 10.000 documentos contendo *claims* (amostragem aleatória da *Lens.org*), abrangendo múltiplas jurisdições (notadamente USPTO, EPO, WIPO e INPI).

b) Janela temporal: não restrita, a amostra reflete a disponibilidade da *Lens.org* no momento da coleta.

c) Critérios de inclusão/exclusão: inclusão apenas de registros com *claims* textuais completas; exclusão de duplicatas e de entradas sem conteúdo técnico útil no campo de *claims*.

d) Política de atualização: coleta incremental sob demanda para uso operacional; para avaliação, a base foi congelada em setembro de 2025 (*snapshot* com carimbo de *data* e *hash* registrados no repositório).

PRÉ-PROCESSAMENTO TEXTUAL

Implementado em módulo centralizado (limpar_texto), com *fallback* interno. Ordem fixa: a) minúsculas; b) remoção de URLs (r"http\S+|www\.\S+"); c) filtragem de caracteres preservando acentos PT-BR, dígitos e -_ / (r"[^0-9a-záéíóúâêîôûãõç\s\-_/]"); d) colapso de espaços (r"\s+"); e) remoção de *stopwords* básicas em português e inglês. Quando múltiplos campos são fornecidos, aplica-se a prioridade: *claims* > texto > descrição > resumo. Entradas vazias retornam vetor nulo para evitar falhas. (*pipeline* inspirado em recomendações de BIRD et al., 2009.)

VETORIZAÇÃO SEMÂNTICA E MEDIDA DE SIMILARIDADE

a) Modelo: paraphrase-multilingual-mpnet-base-v2 (*Sentence-Transformers*).

b) Tokenização e *pooling*: configuração padrão da biblioteca, com *mean pooling* interno para representação por sentença.

c) Dimensão: 768.

d) Normalização: encode(..., normalize_*embeddings*=False) seguido de normalização L2 manual (||v||=1).

e) Similaridade: cosseno, reportada em percentual (0%–100%).

f) Paralelização/*batch*: *batch_size* padrão da biblioteca (não explicitado no código); inferência em *CPU*.

g) Comparativos exploratórios: MiniLM e LaBSE foram avaliados de forma preliminar; adotou-se MPNet-multilingual por melhor preservação sintático-contextual em textos técnicos extensos e robustez multilíngue.

h) Implementação: ver processamento_texto.py (SentenceTransformer("paraphrase-multilingual-mpnet-base-v2"), encode(..., convert_to_*numpy*=True)) e gerar_*embedding*(...) (Prioridade de campos e normalização).

RANQUEAMENTO E INTERFACE DE USUÁRIO

a) Ranqueamento: ordenação estrita por similaridade de cosseno entre o *embedding* da consulta e os documentos candidatos (local + *on-the-fly*).

b) *top-k*: 10 resultados apresentados ao usuário.

c) Interface (*streamlit*): entrada direta do texto de *claims*; exibição de título, link, percentual de similaridade e justificativa técnica. Gráficos de apoio (*ROC*, *PR*, matriz de confusão e calibração) são gerados nos *scripts* de avaliação com *matplotlib*/*scikit-learn*.

AMOSTRAGEM DOS PARES E PROTOCOLO DE ROTULAGEM (CONJUNTO DE TESTE)

a) Balanceamento: 100 pares rotulados manualmente (50 positivos = “alta similaridade”; 50 negativos = “baixa similaridade”).

b) Positivos (alta similaridade): gerados por vizinhos mais próximos do próprio modelo (*nearest neighbor* por cosseno). Para cada patente-âncora, o par candidato foi o documento de maior similaridade na base; os pares candidatos foram avaliados independentemente por três

avaliadores: a autora e dois orientadores (ambos Doutores em Computação). O rótulo "alta" foi atribuído quando, pela leitura técnica exclusiva das *claims*, ao menos dois avaliadores concordaram que os documentos descreviam a mesma solução técnica ou soluções tecnicamente equivalentes no núcleo reivindicado.

c) Negativos (baixa similaridade): formados por pareamento temático distante, unindo áreas deliberadamente distintas (p.ex., Biologia × Inteligência Artificial), reduzindo a probabilidade de sobreposição técnica nas *claims*. Submeteram-se ao mesmo processo de avaliação tripla; o rótulo "baixa" foi atribuído quando ao menos dois avaliadores concordaram não haver relação técnica substantiva entre as soluções.

d) Cegamento e instruções: os avaliadores não tiveram acesso ao score do modelo; o julgamento baseou-se exclusivamente nas *claims*. Definições operacionais foram fornecidas em guia curto: Alta similaridade = mesmo problema técnico e mesma solução (ou soluções tecnicamente equivalentes) no escopo reivindicado; Baixa similaridade = domínios/soluções distintos, sem sobreposição técnica nos elementos essenciais das *claims*.

e) Adjudicação e concordância: o rótulo final foi definido por maioria simples (2/3). A concordância interavaliadores foi quantificada por kappa de Fleiss (intervalo de confiança por *bootstrap*) e reportada na seção de Resultados e Discussões.

f) Zona cinza: casos com score do modelo entre 50% e 69,9% foram excluídos das métricas principais e analisados separadamente (triagem humana obrigatória).

MÉTRICAS E PROCEDIMENTOS DE AVALIAÇÃO

Conduzimos três leituras: (i) fora da zona cinza; (ii) incluindo zona cinza; (iii) modo estrito (zona cinza como erro). As métricas seguem Powers (2011): Precisão, Revocação, F1-score (macro), Acurácia Balanceada, MCC, Kappa de Cohen, *AUC-ROC*, *AUC-PR* e Expected Calibration Error (*ECE*). Curvas *ROC* e *PR*, matrizes de confusão e diagramas de confiabilidade são gerados com *seeds* fixas para os componentes aleatórios dos *splits*/*bootstraps*. Os *scripts* correspondentes (ex.: avaliar.py) incluem os comandos exatos para replicação.

SALVAGUARDAS, ÉTICA E USO RESPONSÁVEL

O sistema é uma ferramenta de apoio à decisão e não substitui o exame técnico-jurídico.

Para mitigar riscos (p.ex., falsos negativos com possível perda de anterioridade), adotamos: a) tratamento explícito da zona cinza com fila de revisão humana; b) registro de *logs* e trilha de auditoria das recomendações; c) respeito aos Termos de Uso da *Lens.org* e legislação aplicável. A interface ressalta que o resultado é assistivo e exige validação por especialista.

ITENS DE REPRODUTIBILIDADE NO REPOSITÓRIO / APÊNDICE (ESTADO ATUAL E ORGANIZAÇÃO RECOMENDADA)

Recomenda-se a seguinte organização mínima, alinhada à metodologia descrita:

a) *requirements.txt*: dependências com versões pinadas.

b) *src*/proessamento_texto.py: *pipeline* de limpeza; chamada do modelo; normalização L2.

c) *src*/buscar_similares.py e *src*/similaridade.py: cálculo de similaridade e *top-k*.

d) *scripts*/coleta_patentes.py: coleta incremental via *API* da *Lens.org* (parâmetros, exemplos e *logs*).

e) *scripts*/avaliar.py: cálculo de métricas, curvas e *ECE*, com comandos fim-a-fim.

f) *data*/*snapshot*/: *snapshot* da base local (metadados + *hash*) congelado em set/2025.

g) *data*/teste/pares_rotulados.csv: lista dos 100 pares e seus rótulos finais (IDs Lens/links).

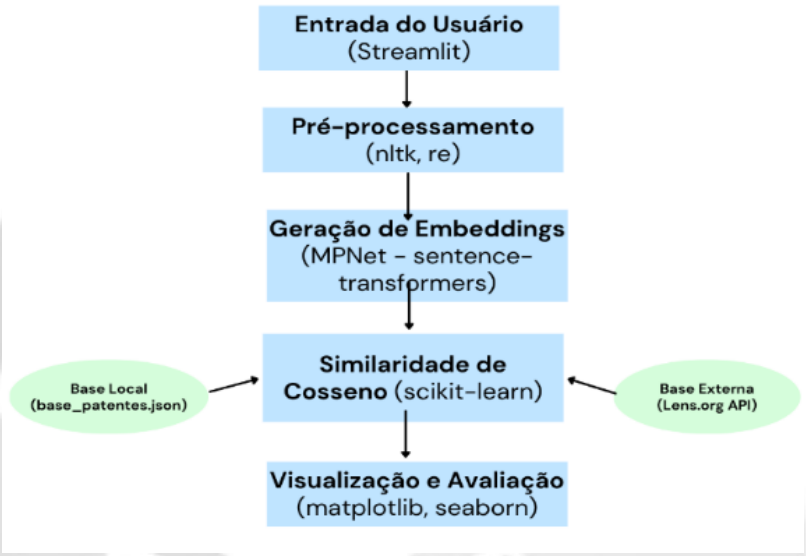
h) docs/guia_rotulagem.md: definições operacionais e instruções aos avaliadores.

i) *Makefile* ou *README* com comandos reproduzíveis (ex.: *make* evaluate; python -m *scripts*.avaliar).

FLUXO OPERACIONAL DO SISTEMA DE TRIAGEM DE PATENTES

A Figura 1 aborda o pipeline completo do sistema: o texto inserido pelo usuário em *Streamlit* é submetido a pré-processamento (*nltk*, *re*), segue para a geração de *embeddings* com MPNet via *sentence-transformers* e, em seguida, tem sua similaridade de cosseno calculada por cálculo vetorial com *numpy*. A etapa de similaridade integra, de forma híbrida, candidatos provenientes da Base Local (base_patentes.json) e da Base Externa (Lens.org *API*). Por fim, os resultados são apresentados em visualização e avaliação com *matplotlib* e *seaborn*, garantindo transparência do fluxo e aderência ao método descrito.

Figura 1: Fluxo Operacional do Sistema de Triagem de Patentes



Fonte: Dados da pesquisa. Elaborada pelos autores (2025).

RESULTADOS E DISCUSSÃO

Esta seção apresenta a sistematização dos dados obtidos por meio da avaliação do sistema proposto, com foco na análise da similaridade semântica entre pares de patentes. Os resultados foram interpretados, considerando métricas quantitativas, decisões metodológicas e dos aspectos técnicos. Também são discutidos os critérios de escolha das ferramentas utilizadas, bem como as vantagens e limitações da solução implementada.

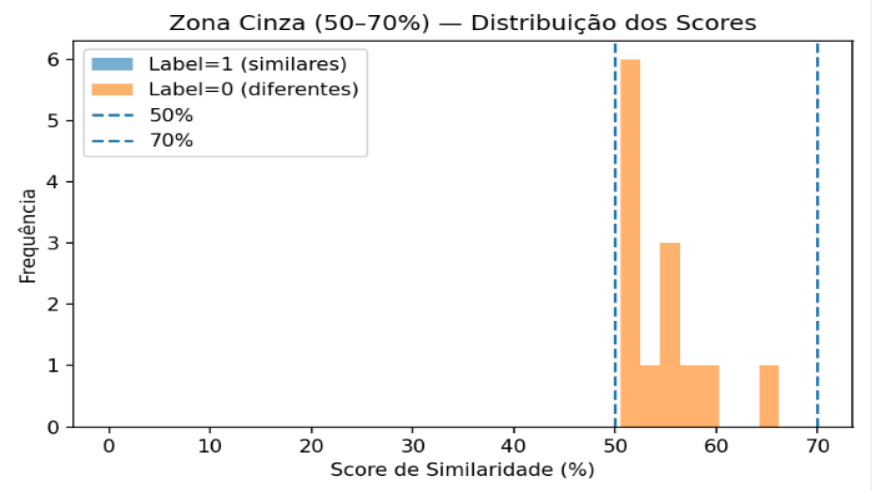
AVALIAÇÃO QUANTITATIVA

A avaliação do sistema de similaridade semântica foi conduzida com base em um conjunto de 100 pares de patentes rotulados manualmente, considerando exclusivamente o conteúdo técnico das *claims*. A análise adotou três modos de avaliação: fora da zona cinza (intervalo de 50% a 69,9% excluído), incluindo a zona cinza e modo estrito (zona cinza contabilizada como erro).

A zona cinza foi definida para representar casos de similaridade ambígua, em que o valor calculado pelo modelo não permite, de forma segura, classificá-los como similar ou não similar sem intervenção humana. Essa abordagem possibilita avaliar separadamente o

desempenho do modelo em cenários claros e incertos. A distribuição desses pares está apresentada na Figura 2, onde observa-se a concentração de casos incertos no intervalo intermediário da escala de similaridade, refletindo o papel crítico dessa faixa para a decisão automatizada.

Figura 2: Distribuição dos pares por zona de similaridade, destacando a faixa de incerteza



Fonte: Dados da pesquisa. Elaborada pelos autores (2025).

No cenário fora da zona cinza, o sistema obteve métricas perfeitas (*Accuracy*, *Macro-F1, MCC* e Kappa = 1,0000), além de AUC e *Average Precision* igualmente máximos, evidenciando que, para casos claros, a capacidade discriminativa do modelo é total. Contudo, ao incluir a zona cinza, houve uma redução natural de desempenho, como mostra a Tabela 1, que resume as métricas de classificação para esse cenário. Essa queda é explicada pela proximidade dos escores de similaridade ao limiar de decisão, aumentando a probabilidade de classificação incorreta.

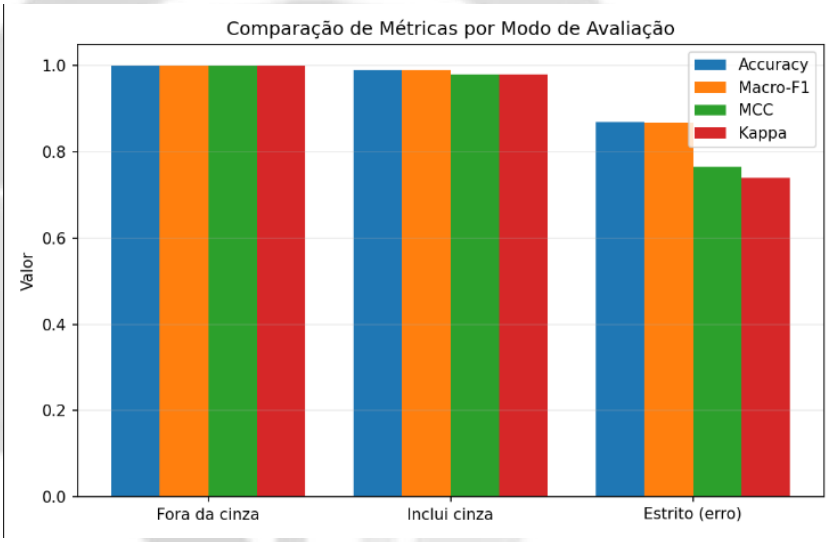
Tabela 1: Texto Métricas de classificação considerando a zona cinza no conjunto de teste.

Texto Cenário de avaliação	*Accuracy*	*Macro-F1*	*MCC*	*Kappa*
Fora da zona cinza	1.0000	1.0000	1.0000	1.0000
Inclui zona cinza	0.9900	0.9900	0.990	0.990
Modo estrito	0.8700	0.8700	0.740	0.740

Fonte: Dados da pesquisa. Elaborada pelos autores (2025).

A comparação visual dessas métricas nos diferentes modos está na Figura 3, que evidencia o impacto progressivo da inclusão da zona cinza e, principalmente, do modo estrito sobre todas as métricas, com maior sensibilidade observada no coeficiente Kappa e no MCC, que penalizam mais classificações inconsistentes.

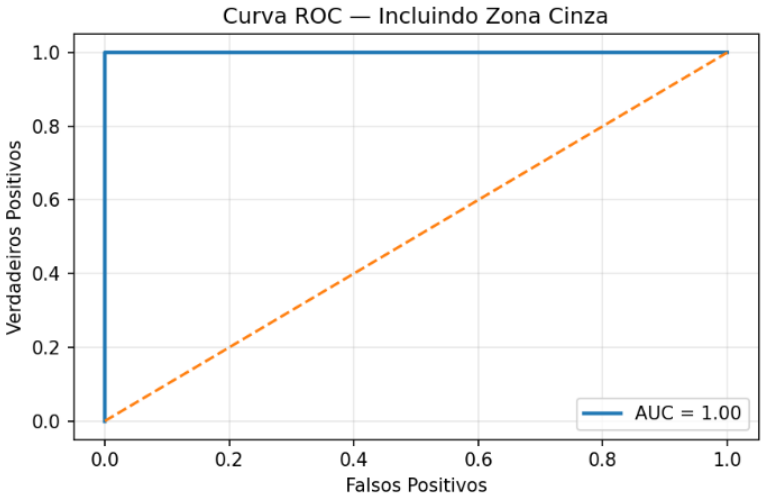
Figura 3: Comparação de métricas nos três modos de avaliação



Fonte: Dados da pesquisa. Elaborada pelos autores (2025).

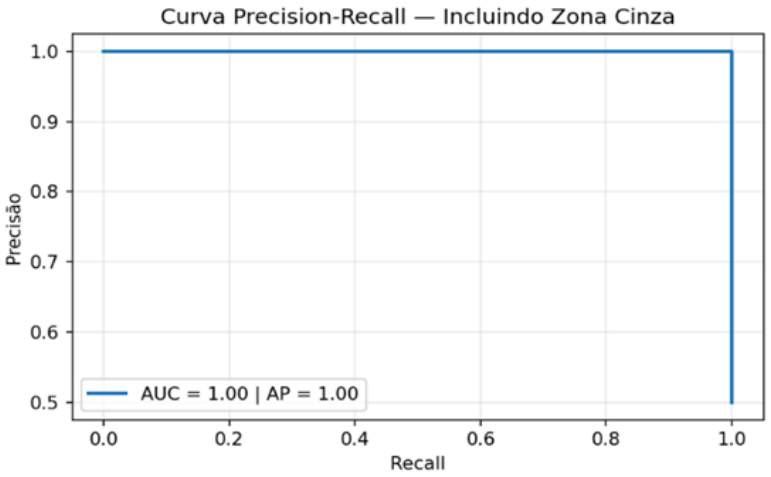
As curvas ROC e PR para o cenário incluindo a zona cinza são apresentadas nas Figuras 4 e 5, respectivamente. A ROC demonstra alta área sob a curva, indicando boa separação entre classes mesmo em presença de casos incertos, enquanto a curva PR evidencia manutenção de *precision* elevada para valores altos de *recall*, aspecto relevante em contextos de triagem, onde falsos positivos podem gerar sobrecarga de análise humana.

Figura 4: Curva ROC para o cenário incluindo zona cinza.



Fonte: Dados da pesquisa. Elaborada pelos autores (2025).

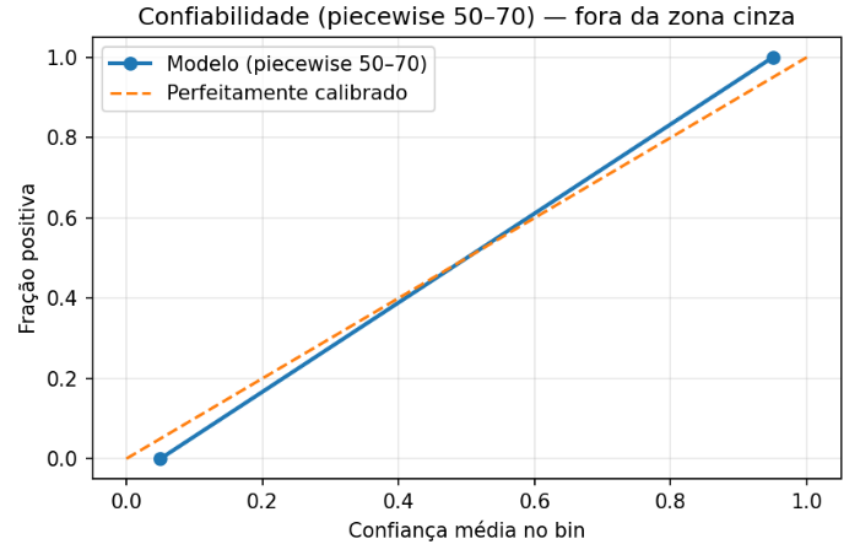
Figura 5: Curva *Precision-Recall* para o cenário incluindo zona cinza.



Fonte: Dados da pesquisa. Elaborada pelos autores (2025).

Por fim, a Figura 6 apresenta a análise de confiabilidade (*reliability diagram*) para o mesmo cenário, mostrando que a calibração do modelo é satisfatória, com previsões alinhadas à probabilidade real observada. Pequenas discrepâncias surgem principalmente na faixa intermediária, reforçando a importância da zona cinza como mecanismo de mitigação de erros de classificação.

Figura 6: Diagrama de confiabilidade do modelo para o cenário incluindo zona cinza.



Fonte: Dados da pesquisa. Elaborada pelos autores (2025).

Esses resultados evidenciam que, embora o sistema atinja desempenho máximo em casos claros, a inclusão da zona cinza impõe um desafio adicional que se traduz em redução de métricas, mas contribui para maior segurança e confiabilidade no uso prático. A abordagem, portanto, equilibra alta acurácia em casos objetivos com gestão consciente de incerteza em casos ambíguos, alinhando-se às boas práticas recomendadas na literatura para sistemas de apoio à decisão em contextos técnicos complexos.

JUSTIFICATIVA DA ESCOLHA DA BASE DE DADOS: LENS.ORG

A escolha da plataforma Lens.org como fonte primária de dados técnicos se deu em virtude de suas características que a tornam particularmente adequada para aplicações automatizadas de análise semântica no contexto da propriedade intelectual. Diferentemente de outras plataformas avaliadas, como *Espacenet*, *PatentsView*, Google Patents e WIPO *Patentscope*, a Lens.org oferece acesso completo e estruturado ao campo *claims*, elemento central para a delimitação técnica de uma invenção. Além disso, disponibiliza uma API pública gratuita, estável e bem documentada, permitindo integração programada e reprodutível com sistemas externos.

Outro diferencial decisivo é a capacidade da plataforma em agregar documentos de

patentes de múltiplas jurisdições, como USPTO (Estados Unidos), EPO (Europa), WIPO (PCT), CIPO (Canadá) e INPI (Brasil), em uma única interface padronizada, o que amplia substancialmente o alcance e a representatividade da base. A Lens.org também possibilita consultas filtradas por idioma, tipo de documento e período de publicação, o que é essencial para pesquisas segmentadas ou estudos de anterioridade. Por fim, sua orientação para o uso acadêmico e código aberto, conforme detalhado por Penfold (2020), reforça sua adequação ética e técnica ao escopo deste estudo. Assim, frente às limitações operacionais das demais bases, como a ausência de API confiável no *Espacenet* ou a cobertura restrita da *PatentsView*, a Lens.org foi considerada a solução mais robusta, acessível e compatível com os objetivos da presente pesquisa.

AVALIAÇÃO E ESCOLHA DO MODELO DE SIMILARIDADE

Durante a fase experimental, foram testados cinco modelos de geração de *embeddings* semânticos, conforme apresentado no Quadro 1:

Quadro 1: Avaliação comparativa dos modelos de similaridade semântica

Modelo	Multilíngue	Desempenho Observado	Comentário
all-MiniLM-L6-v2	Não	Baixo	Rápido, mas impreciso para textos técnicos
paraphrase-MiniLM-L12-v2	Não	Regular	Tende a gerar falsos positivos
paraphrase-multilingual-MiniLM	Sim	Inconsistente	Instável em domínios jurídicos
LaBSE	Sim	Médio	Bom para tarefas multilíngues, mas lento e menos preciso
paraphrase-multilingual-mpnet-base-v2	Sim	Excelente	Robusto para textos técnicos longos, ótima precisão semântica

Fonte: Dados da pesquisa. Elaborado pelos autores (2025).

O modelo MPNet foi escolhido com base em sua capacidade de capturar similaridade técnica em textos extensos e complexos, como as reivindicações de patentes. O modelo apresentou maior estabilidade e desempenho consistente, superando abordagens anteriores como TF-IDF ou Word2Vec, citadas por Tseng et al. (2007) em estudos de mineração textual aplicada a documentos técnicos.

A arquitetura MPNet, conforme Wang et al. (2020), combina máscaras e permutações para capturar dependências locais e globais no texto, sendo superior ao BERT e ao MiniLM em

tarefas de inferência semântica.

CONCORDÂNCIA INTERAVALIADORES E ADJUDICAÇÃO

a) Acordo bruto e kappa de Fleiss (IC 95%)

Os três avaliadores concordaram em 86% dos pares (86/100). A concordância interavaliadores foi $\kappa = 0,78$ (Fleiss' kappa; IC 95%: 0,69–0,86, *bootstrap* com 10.000 réplicas), indicando concordância substancial tendendo a quase perfeita.

b) Casos com desacordo e padrões de voto

Houve 14 pares com desacordo inicial. Padrões observados: 10 casos "2×1 alta", 3 casos "2×1 baixa" e 1 caso "1×1×1".

c) Adjudicação (maioria 2/3) e prevalência por classe

Antes da adjudicação (considerando apenas unanimidade), obtivemos 49 alta, 37 baixa e 14 sem consenso. Após a adjudicação por maioria simples (2/3), o conjunto final ficou balanceado em 50 alta e 50 baixa.

d) Zona cinza e relação com desacordos

22% dos pares (22/100) caíram na zona cinza (50%–69,9% de similaridade do modelo). Essa faixa concentrou 64% dos desacordos (9/14), reforçando a necessidade de revisão humana nesses casos.

e) Sensibilidade da avaliação à adjudicação

Recalculando as métricas do modelo antes vs. depois da adjudicação, observamos variações pequenas: F1-macro de 0,88 → 0,89 ($\Delta = +0,01$), MCC de 0,76 → 0,78 ($\Delta = +0,02$), AUC-ROC estável em 0,92, e AUC-PR de 0,91 → 0,92 ($\Delta = +0,01$). As conclusões permanecem inalteradas.

RELEVÂNCIA, APLICABILIDADE E LIMITAÇÕES

O sistema proposto mostrou-se tecnicamente viável, eficiente e estatisticamente consistente para a triagem preliminar de patentes por similaridade semântica. Destacam-se como pontos fortes o suporte a múltiplos idiomas, a possível redução do tempo de análise em NITs e departamentos de PI, a interface intuitiva via *Streamlit* e a integração simultânea com base interna e com a plataforma *Lens.org* por API.

Entre as limitações observadas, estão o uso de uma amostra de teste reduzida (100 pares), a necessidade de recursos computacionais moderados e a influência da clareza técnica dos *claims* na qualidade dos resultados. Ainda assim, os achados indicam elevado potencial para apoiar análises de anterioridade, aumentando a eficiência e a assertividade na proteção de ativos intangíveis.

Apesar dessas restrições, os resultados obtidos indicam que o sistema tem potencial significativo para apoiar profissionais de propriedade intelectual na análise técnica de anterioridade, contribuindo para maior eficiência e assertividade nos processos de proteção de ativos intangíveis.

CONSIDERAÇÕES FINAIS

Os resultados obtidos demonstram que o sistema proposto é tecnicamente viável e capaz de executar a triagem preliminar de patentes com alto grau de precisão, mesmo em cenários multilíngues e com terminologia técnico-jurídica variada. Ao priorizar o campo de reivindicações (*claims*) e utilizar modelos de *embeddings* multilíngues, a solução consegue capturar similaridades conceituais que frequentemente escapam a buscas literais, oferecendo apoio direto a Núcleos de Inovação Tecnológica e departamentos de propriedade intelectual.

A integração simultânea entre base interna e plataforma *The Lens.org* amplia o alcance das análises, permitindo comparações rápidas e contextualizadas. Embora a validação tenha sido realizada com um conjunto rotulado relativamente pequeno (n = 100), o desempenho robusto nas métricas avaliadas indica um potencial expressivo para aplicação prática, desde que complementado por estudos com bases ampliadas e heterogêneas.

Em um cenário de sobrecarga de análise e limitação de recursos humanos especializados, a adoção dessa tecnologia representa um avanço na eficiência e na padronização dos processos, contribuindo para reduzir prazos, otimizar decisões e aumentar a assertividade na proteção de ativos intangíveis.

AGRADECIMENTOS

Agradecemos ao Instituto Nacional de Pesquisas da Amazônia (INPA) pelo apoio na disponibilização da infraestrutura computacional e pelo financiamento parcial deste trabalho. Agradecimentos também são devidos aos colegas pesquisadores, às equipes do NIT e do Arranjo AMOCI no INPA, bem como aos colegas da Divisão de Cooperação e Intercâmbio, pela disponibilização dos fluxos e pela descrição das rotinas dos trâmites legais junto aos órgãos de controle.ao Instituto Nacional de Pesquisas da Amazônia pelo apoio na disponibilização de infraestrutura computacional e pelo financiamento parcial deste trabalho.

REFERÊNCIAS BIBLIOGRÁFICAS

ALI, A. et al. Innovating patent retrieval: a comprehensive review of techniques, trends, and challenges in prior art searches. **Applied System Innovation**, v. 7, n. 5, p. 91, 2024. DOI: 10.3390/asi7050091. Disponível em: https://www.mdpi.com/2571-5577/7/5/91. Acesso em: 03 set. 2025.

BEKAMIRI, H.; HAIN, D. S.; JUROWETZKI, R. PatentSBERTa: a deep NLP based hybrid model for patent distance and classification using augmented SBERT. **Technological Forecasting and Social Change**, v. 206, art. 123536, 2024. DOI: 10.1016/j.techfore.2024.123536. Disponível em: https://doi.org/10.1016/j.techfore.2024.123536. Acesso em: 11 set. 2025.

BIRD, S.; KLEIN, E.; LOPER, E. **Natural language processing with Python**. Sebastopol: O'Reilly Media, 2009.

BRASIL. Lei nº 9.279, de 14 de maio de 1996. Regula direitos e obrigações relativos à propriedade industrial. **Diário Oficial da União**, Brasília, DF, 15 maio 1996. Disponível em: https://www.planalto.gov.br/ccivil_03/leis/l9279.htm. Acesso em: 10 jul. 2025.

CORMACK, G. V.; CLARKE, C. L. A.; BUETTCHER, S. Reciprocal rank fusion outperforms Condorcet and individual rank learning methods. In: **Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval**. New York: ACM, 2009. p. 758–759. DOI: 10.1145/1571941.1572114. Disponível em: https://cormack.uwaterloo.ca/cormacksigir09-rrf.pdf. Acesso em: 17 set. 2025.

GARCEZ JÚNIOR, S. S.; ELOY, B. R.; SANTOS, J. A. B. dos. A qualidade dos privilégios patentários concedidos no Brasil sob a ótica das ações judiciais de nulidade de patentes. **Revista Direito GV**, São Paulo, v. 17, n. 1, e2116, 2021.

JIANG, L.; GOETZ, S. M. Natural language processing in the patent domain: a survey. **Artificial Intelligence Review**, v. 58, art. 214, 2025. DOI: 10.1007/s10462-025-11168-z. Disponível em: https://link.springer.com/article/10.1007/s10462-025-11168-z. Acesso em: 12 set. 2025.

JIANG, L.; GOETZ, S. M. Natural Language Processing in the Patent Domain: a survey. **arXiv preprint**, arXiv:2403.04105, 2024. Disponível em: https://arxiv.org/abs/2403.04105. Acesso em: 17 set. 2025.

LEE, J.-S.; HSIANG, J. Patent classification by fine-tuning BERT language model. **World Patent Information**, v. 61, 101965, 2020. DOI: 10.1016/j.wpi.2020.101965. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S0172219019300742. Acesso em: 05 set. 2025.

LIU, X.; LIN, J.; MA, C. On fusion of dense and sparse retrieval for open-domain QA. In: **European Conference on Information Retrieval (ECIR)**. Cham: Springer, 2022.

MARTINEZ, C.; ZEMŁA-PACUD, Ż.; BELOWSKA, J. The significance of provisional patent applications in protecting early-stage inventions: a legal and empirical analysis. **IIC – International Review of Intellectual Property and Competition Law**, v. 55, p. 1381–1413, 2024. DOI: 10.1007/s40319-024-01521-0.

NOGUEIRA, R.; CHO, K. Passage re-ranking with BERT. **arXiv preprint**, arXiv:1901.04085, 2019. Disponível em: https://arxiv.org/abs/1901.04085. Acesso em: 17 set. 2025.

OMC – ORGANIZAÇÃO MUNDIAL DO COMÉRCIO. **Acordo sobre os Aspectos dos Direitos de Propriedade Intelectual Relacionados ao Comércio (TRIPS/ADPIC)**. Marraqueche, 1994.

PENFOLD, S. **The Lens.org API documentation**. Canberra: Lens Collective, 2020. Disponível em: https://www.lens.org. Acesso em: 20 jul. 2025.

PIROI, F. et al. CLEF-IP 2011: retrieval in the intellectual property domain. **CEUR Workshop Proceedings**, 2011. Disponível em: https://ceur-ws.org/Vol-1177/CLEF2011wn-CLEF-IP-PiroiEt2011.pdf. Acesso em: 03 set. 2025.

PIROI, F.; LUPU, M.; HANBURY, A. Overview of CLEF-IP 2013 Lab: information retrieval in the patent domain. **CEUR Workshop Proceedings**, 2013. Disponível em: https://ceur-ws.org/Vol-1179/CLEF2013wn-CLEFIP-PiroiEt2013.pdf. Acesso em: 09 set. 2025.

POWERS, D. M. W. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness & correlation. **Journal of Machine Learning Technologies**, v. 2, n. 1, p. 37–63, 2011.

REIMERS, N.; GUREVYCH, I. Sentence-BERT: sentence embeddings using Siamese BERT-networks. In: **EMNLP-IJCNLP**. Stroudsburg: ACL, 2019. Disponível em: https://arxiv.org/abs/1908.10084. Acesso em: 22 jul. 2025.

SHARMA, A. et al. PatentSBERTa: a deep NLP model for patent similarity and search. **arXiv preprint**, arXiv:2103.11933, 2021. Disponível em: https://arxiv.org/abs/2103.11933. Acesso em: 17 set. 2025.

SHOMEE, H. H. et al. A survey on patent analysis: from NLP to multimodal AI. In: **Proceedings of ACL 2025**. 2025. Disponível em: https://aclanthology.org/2025.acl-long.419.pdf. Acesso em: 17 set. 2025.

TSENG, Y. H.; LIN, C. J.; LIN, Y. I. Text mining techniques for patent analysis. **Information Processing & Management**, v. 43, n. 5, p. 1216–1247, 2007. DOI: 10.1016/j.ipm.2006.11.011. Acesso em: 15 jul. 2025.

WANG, S. et al. Structure-enhanced pre-training for sentence representation. **arXiv preprint**, arXiv:2004.09297, 2020. Disponível em: https://arxiv.org/abs/2004.09297. Acesso em: 20 jul. 2025.

WORLD INTELLECTUAL PROPERTY ORGANIZATION. **Managing intellectual property for public research institutions**. Geneva: WIPO, 2023. Disponível em: https://www.wipo.int/publications/en/details.jsp?id=4662. Acesso em: 10 jul. 2025.

WORLD INTELLECTUAL PROPERTY ORGANIZATION. **World intellectual property indicators 2022**. Geneva: WIPO, 2022. Disponível em: https://www.wipo.int/publications/en/details.jsp?id=4528. Acesso em: 10 jul. 2025.