26º ENANCIB
Encontro Nacional de Pesquisa e Pós-Graduação em Ciência da Informação
A Amazônia na sociedade global em rede: assimetrias e potencialidades informacionais, tecnológicas e ambientais
Belém, Pará, 09 a 13 de novembro de 2026
Universidade Federal do Pará

XXVI ENCONTRO NACIONAL DE PESQUISA E PÓS-GRADUAÇÃO EM CIÊNCIA DA INFORMAÇÃO - XXVI ENANCIB

ISSN 2177-3688

GT 2 – Organização e Representação do Conhecimento

INDEXAÇÃO AUTOMÁTICA NA ERA DOS MODELOS DE LINGUAGEM

AUTOMATIC INDEXING IN THE AGE OF LANGUAGE MODELS

Marco Antônio Gusmão Carvalho – Universidade Estadual Paulista "Júlio de mesquita Filho" (Unesp) Campus de Marília
Mariângela Spotti Lopes Fujita – Universidade Estadual Paulista "Júlio de mesquita Filho" (Unesp) Campus de Marília
Isidoro Gil Leiva – Universidad de Murcia (UM)

Modalidade: Trabalho Completo

Resumo: Bases bibliográficas e repositórios institucionais incorporam milhões de documentos anualmente, tornando inviável o tratamento exclusivamente manual. Nesse contexto, a indexação é essencial para a organização e recuperação da informação. Esta pesquisa compara três paradigmas de indexação automática sob vocabulário controlado: um sistema heurístico baseado em regras e reconhecimento terminológico, uma abordagem supervisionada de aprendizado de máquina para classificação multilabel e uma abordagem baseada em modelo de linguagem de grande porte, implementada em um pipeline híbrido de recuperação lexical e semântica. Foi construído um corpus com mais de 9.500 documentos da área de Agricultura, incluindo 8.000 documentos para treinamento e três conjuntos de teste com 500 documentos cada. A avaliação utilizou padrões de referência derivados da base AGRICOLA, obtidos por replicação de estratégias de busca e alinhamento manual das indexações. A indexação foi restrita ao vocabulário controlado da National Agricultural Library, previamente normalizado e representado em formatos textual e vetorial. Os resultados mostram que a comparação entre as abordagens não pode ser reduzida à substituição tecnológica: cada paradigma responde de modo distinto às exigências da indexação automática. No corpus investigado, soluções híbridas que articulam controle terminológico, recuperação textual e inferência semântica mostram-se mais adequadas para explorar o potencial dos modelos de linguagem do que seu uso como indexadores autossuficientes.
Palavras-chave: indexação automática; avaliação da indexação; modelos de linguagem.

Abstract: Bibliographic databases and institutional repositories incorporate millions of documents annually, making exclusively manual processing unfeasible. In this context, indexing is essential for information organization and retrieval. This research compares three

paradigms of automatic indexing under controlled vocabulary: a heuristic system based on rules and terminological recognition, a supervised machine learning approach for multilabel classification, and a large language model approach implemented in a hybrid pipeline combining lexical and semantic retrieval. A corpus of more than 9,500 documents in Agriculture was constructed, including 8,000 documents for training and three test sets with 500 documents each. The evaluation used reference standards derived from the AGRICOLA database, obtained by replicating search strategies and manually aligning the indexing. Indexing was restricted to the controlled vocabulary of the National Agricultural Library, previously normalized and represented in textual and vector formats. The results show that comparison among the approaches cannot be reduced to technological substitution: each paradigm responds differently to the requirements of automatic indexing. In the investigated corpus, hybrid solutions combining terminological control, textual retrieval, and semantic inference appear more suitable for exploring the potential of language models than using them as self-sufficient indexers.
Keywords: automatic indexing; indexing evaluation; language models.

1 INTRODUÇÃO

Desde meados do século XX, a produção documental, seu processamento e sua difusão vêm se expandindo de forma contínua, aumentando a pressão sobre sistemas de informação responsáveis pela organização e pelo acesso aos registros. Esse crescimento pode ser dimensionado por indicadores recentes: o World Bank Open Data registrou, em 2021, mais de três milhões de artigos científico-técnicos, enquanto os dados da rede espanhola de bibliotecas universitárias evidenciam expansão persistente das coleções eletrônicas sem crescimento proporcional das equipes responsáveis por seu tratamento (World Bank, 2021; REBIUN, 2023). Em contexto institucional, Poley *et al*. (2025) registram que a Biblioteca Nacional da Alemanha incorporou, somente em 2023, cerca de 2,6 milhões de publicações digitais. Esse aumento do fluxo documental torna progressivamente oneroso o tratamento intelectual exclusivamente manual e reforça a necessidade de métodos capazes de ampliar a escala da organização e da recuperação da informação.

O desenvolvimento de programas voltados à automação da indexação documental iniciou-se no final da década de 1950 e evoluiu significativamente desde os anos 1970, com sistemas como SMART, CLARIT, SIMPR e SHAPIRE, que integraram análises linguísticas e estatísticas (Gil-Leiva, 2008; Gil-Leiva; Díaz Ortuño; Corrêa, 2022). Mais recentemente, bibliotecas nacionais adotaram soluções baseadas em aprendizado de máquina, como o Annif, da Biblioteca Nacional da Finlândia (Suominen, 2019; Suominen; Inkinen; Lehtinen, 2022), o projeto Kratt, da Estônia (Asula *et al*., 2021), e o projeto EMa, da Biblioteca Nacional da

Alemanha, voltado à atribuição automática de categorias da Classificação Decimal de Dewey (CDD) e descritores temáticos (Poley *et al*., 2025). Essa evolução foi ampliada atualmente com modelos *transformer*, como o *Bidirectional Encoder Representations from Transformers* (BERT), e *Large Language Models* (LLMs) como ChatGPT, Gemini e Copilot.

Os LLMs ampliaram a capacidade de interpretar linguagem natural, inferir relações semânticas e produzir respostas contextualizadas, reacendendo o interesse por sua aplicação em tarefas documentárias. Contudo, compreender semanticamente um documento não implica necessariamente representá-lo de forma documentalmente consistente (Dermentzi; Bryant; Rovigo; García-González, 2025). Persiste, assim, a tensão entre inferência semântica e controle terminológico.

Nesse contexto, avaliar modelos generativos apenas por sua fluência textual é insuficiente (Włodarczyk, 2025). A qualidade da indexação depende também da aderência ao vocabulário controlado, da consistência dos descritores, da estabilidade da representação e de sua utilidade prática em processos de indexação assistida.

Neste estudo, considera-se "documentalmente consistente" a indexação que atende simultaneamente a três condições: (a) utiliza descritores autorizados pelo vocabulário controlado adotado; (b) mantém correspondência temática com os assuntos centrais do documento, tendo como referência a indexação profissional do gold standard; e (c) evita variantes livres ou termos externos ao controle terminológico. Operacionalmente, essa consistência é observada pela aderência ao vocabulário e pelo desempenho em precisão, revocação e Medida-F na comparação com o padrão de referência.

A pergunta que orienta a pesquisa pode ser formulada nos seguintes termos: em que medida abordagens baseadas em modelos de linguagem de grande porte podem operar de forma documentalmente consistente em tarefas de indexação sob vocabulário controlado, em comparação com sistemas mais tradicionais de indexação automática? Ao responder a essa questão, o trabalho procura contribuir para a reflexão contemporânea sobre o lugar da inteligência artificial generativa no tratamento temático da informação, deslocando o debate da substituição tecnológica para a articulação entre representação documental, inferência semântica e controle terminológico.

É justamente nesse ponto que este trabalho se insere. O objetivo da pesquisa é comparar três paradigmas de indexação automática sob vocabulário controlado: um sistema heurístico baseado em regras e reconhecimento terminológico, Sistema de Indización Semi-

Automatica (SISA), uma plataforma supervisionada baseada em aprendizado de máquina e classificação multilabel (Annif) e uma abordagem baseada em modelo de linguagem de grande porte (Llama 3), implementada em um *pipeline* híbrido de recuperação lexical e semântica. Mais do que identificar qual sistema apresenta melhor desempenho absoluto, busca-se compreender em que medida cada paradigma se mostra documentalmente adequado à tarefa de indexação automática.

2 INDEXAÇÃO AUTOMÁTICA E MODELOS DE LINGUAGEM

A indexação automática de assuntos constitui um tema central da Organização do Conhecimento, especialmente em sistemas que utilizam vocabulários controlados. Avanços recentes em aprendizado de máquina e modelos de linguagem reconfiguraram esse campo (Chun *et al*., 2025; Liu *et al*., 2025).

Entre os sistemas especializados, o Annif destaca-se por integrar classificação *multilabel*, recuperação lexical e estratégias híbridas. Poley *et al*. (2025) demonstram sua aplicação em fluxos institucionais reais, enquanto Suominen, Inkinen e Lehtinen (2025) evidenciam que arquiteturas híbridas apresentam melhores resultados do que o uso isolado de LLMs.

A comparação entre modelos generativos e sistemas especializados revela resultados competitivos. Włodarczyk (2025), ao comparar o GPT-4o ao DESCRIPTOR, observou desempenho semelhante em precisão, revocação e F1, embora o GPT-4o tenha mostrado vantagem na ordenação de termos relevantes. O estudo reforça a necessidade de converter compreensão semântica em descritores compatíveis com vocabulários controlados.

Luo e Liu (2025) destacam a importância da ancoragem terminológica, enquanto Dermentzi *et al*. (2025) apontam limitações de abordagens puramente generativas. Ainda assim, LLMs demonstram potencial na geração de termos semanticamente ricos e em tarefas como extração automática de termos (Bouzid e Piron, 2024; Chun *et al*., 2025; Liu *et al*., 2025).

3 PROCEDIMENTOS METODOLÓGICOS

Este estudo caracteriza-se como uma pesquisa experimental, de natureza aplicada, com abordagem quantitativa e delineamento comparativo, voltado para a análise do desempenho de diferentes métodos de indexação automática por atribuição em documentos científicos.

O desenho metodológico foi estruturado de modo a garantir a comparabilidade entre os sistemas SISA, Annif e uma arquitetura baseada em LLMs, preservando a equivalência do

corpus, do vocabulário controlado e dos critérios de avaliação. Dessa forma, buscou-se analisar não apenas o desempenho absoluto de cada abordagem, mas também as implicações de seus diferentes pressupostos teóricos — heurístico, estatístico e generativo — na qualidade da representação temática documental.

3.1 CORPUS DOCUMENTAL E GOLD STANDARD

A coleção de Agricultura deriva do conjunto de dados descrito por Gil-Leiva (2025), construído a partir de metadados obtidos na Scopus. A coleta inicial foi realizada por buscas pelo ISSN ou pelo nome de oito periódicos científicos de Agricultura em inglês, abrangendo artigos publicados entre 1977 e 2024 e totalizando mais de 17 mil registros. Para cada artigo foram extraídos título, resumo, palavras-chave de autoria e referências. Do material coletado, foram retidos os registros que puderam ser alinhados ao padrão ouro da AGRICOLA, resultando em mais de 9.500 documentos. O conjunto final foi dividido em 8.000 documentos para treinamento e três subcorpora de teste (C1, C2 e C3), com 500 documentos cada. O dataset utilizado está disponível no repositório Zenodo (Gil-Leiva, 2025).

Os padrões de referência (gold standards) foram obtidos da AGRICOLA, base bibliográfica da National Agricultural Library (NAL), cuja indexação de assuntos utiliza o NAL Agricultural Thesaurus. O alinhamento entre os registros coletados na Scopus e os registros correspondentes da AGRICOLA foi realizado por script em Python que comparava os títulos e, quando identificada a correspondência, separava os metadados do documento e os descritores atribuídos profissionalmente em arquivos pareados. A etapa de consolidação incluiu conferência dos casos de correspondência não unívoca. Os descritores provenientes da AGRICOLA foram então utilizados como gold standard para a avaliação dos três sistemas.

3.2 VOCABULÁRIO CONTROLADO E PADRONIZAÇÃO TERMINOLÓGICA

A pesquisa adotou a indexação sob restrição terminológica, utilizando como referência o NAL Agricultural Thesaurus, versão de 2017, obtido da National Agricultural Library (United States Department of Agriculture, 2017). Para a arquitetura baseada em LLM, o vocabulário operacional foi consolidado a partir dos descritores autorizados associados à coleção e aos padrões de referência, com remoção de duplicidades e normalização para comparação sem distinção entre maiúsculas e minúsculas. Na recuperação lexical, os descritores foram tokenizados em letras minúsculas, com remoção de stopwords, e indexados pelo BM25

(BM25Okapi). Na recuperação semântica, os mesmos descritores foram convertidos em embeddings densos pelo modelo mxbai-embed-large, executado via Ollama; os vetores foram normalizados pela norma L2 e armazenados em um índice FAISS do tipo IndexFlatIP, permitindo busca por similaridade de cosseno mediante produto interno. Os resultados lexical e semântico foram combinados por Reciprocal Rank Fusion (RRF). Quanto ao mapeamento terminológico, a versão revisada do pipeline não aceita livremente um termo gerado pela LLM e depois procura um equivalente no tesauro: a equivalência semântica é utilizada apenas para recuperar candidatos já autorizados. O Llama 3 recebe exclusivamente essa lista de candidatos e sua saída é validada por correspondência exata, sem distinção de caixa, com os descritores fornecidos; qualquer termo externo à lista é descartado. Dessa forma, o controle terminológico é imposto antes e depois do reranqueamento pela LLM.

3.3 PROCEDIMENTOS EXPERIMENTAIS

O SISA foi utilizado como *baseline* heurístico, representando uma abordagem clássica de indexação automática baseada em regras e vocabulário controlado. Seu funcionamento consiste na identificação de termos no texto e na aplicação de heurísticas relacionadas à estrutura documental, sem necessidade de treinamento prévio. Os documentos foram preparados em formato compatível com o sistema, e os descritores foram extraídos diretamente da saída da ferramenta. A escolha do SISA justifica-se por sua proximidade à lógica da indexação humana em contextos de terminologia estável e documentos estruturados.

O Annif foi empregado como *baseline* supervisionada, representando o paradigma contemporâneo de indexação automática baseado em aprendizado de máquina. Sua arquitetura fundamenta-se em algoritmos de classificação *multilabel* com vocabulários extensos, capazes de aprender associações entre documentos e descritores a partir de dados previamente indexados.

O procedimento experimental envolveu a preparação de um conjunto de treinamento composto por pares documento-descritores, a partir do qual o modelo foi ajustado. Após o treinamento, os documentos do conjunto de teste foram submetidos ao sistema para geração de descritores candidatos.

Sua inclusão no experimento permite representar o estado da arte em indexação automática baseada em aprendizado de máquina, especialmente em cenários de grande escala e vocabulários extensos.

A abordagem baseada em LLM foi implementada com o modelo Llama 3 (llama3:8b), executado localmente via Ollama. O documento é representado por seus metadados textuais e submetido a duas estratégias de recuperação de descritores: BM25, para correspondência lexical, e FAISS com embeddings mxbai-embed-large, para similaridade semântica. Por padrão, são recuperados até 30 candidatos em cada estratégia; as listas são fundidas por RRF (k=60) e os 35 candidatos mais bem posicionados são enviados ao Llama 3. O modelo opera com temperatura 0 e é instruído a selecionar somente descritores presentes na lista, retornando no máximo 20 termos em formato JSON; a saída passa ainda por validação estrita contra a própria lista de candidatos. O código-fonte, as dependências, os parâmetros padrão e os comandos para reprodução dos experimentos estão disponíveis publicamente[1]. O repositório também disponibiliza rotinas para preparação e cache do índice, execução independente dos subcorpora C1, C2 e C3 e geração dos relatórios de avaliação.

3.4 ESTRATÉGIA DE COMPARAÇÃO E MÉTRICAS DE AVALIAÇÃO

A comparação entre os sistemas foi realizada com base em um princípio unificador: todos deveriam produzir descritores do mesmo vocabulário controlado, permitindo tratá-los como soluções concorrentes para o mesmo problema informacional. A avaliação utilizou métricas clássicas de recuperação da informação e classificação multilabel, como precisão, revocação e F-medida.

4 RESULTADOS E DISCUSSÃO

Os resultados são apresentados em três níveis complementares. Inicialmente, compara-se o desempenho quantitativo dos sistemas nos três subcorpora; em seguida, interpretam-se as diferenças entre os paradigmas heurístico, supervisionado e generativo; por fim, discute-se especificamente o papel dos LLMs à luz da literatura recente sobre indexação automática sob vocabulário controlado.

4.1 DESEMPENHO COMPARATIVO ENTRE SISA, ANNIF E ARQUITETURA BASEADA EM LLMS

A análise comparativa dos resultados evidencia diferenças relevantes entre os três paradigmas de indexação automática considerados nesta pesquisa: a abordagem heurística baseada em regras (SISA), a abordagem supervisionada baseada em aprendizado de máquina

[1] Disponível em: https://github.com/marcoweb/llama3-first-experiment

(Annif) e a abordagem baseada em modelo de linguagem de grande porte integrada a um pipeline híbrido de recuperação lexical e semântica (LLM).

De modo geral, os dados mostram que os melhores desempenhos globais foram obtidos por abordagens tradicionais e supervisionadas, especialmente no corpus de Agricultura, enquanto a abordagem baseada em LLMs apresentou desempenho inferior em termos de precisão, revocação e F-medida. Ainda assim, a comparação revela aspectos importantes sobre o papel de cada paradigma na indexação documental sob vocabulário controlado.

Tabela 1 – Comparação entre paradigmas de indexação automática por domínio e cenário

Sistema	Variante	Corpus	Precisão	Revocação	F-Measure
SISA	—	C1	48,15	49,18	47,67
Annif	MLLM	C1	74,26	40,38	52,31
Annif	Ensemble	C1	64,30	34,66	45,04
LLM	Pipeline híbrido	C1	33,28	17,82	23,21
SISA	—	C2	57,75	74,43	65,04
Annif	MLLM	C2	80,15	50,36	61,86
Annif	Ensemble	C2	66,87	42,22	51,76
LLM	Pipeline híbrido	C2	32,72	20,32	25,07
SISA	—	C3	81,45	51,05	62,76
Annif	MLLM	C3	59,26	69,06	63,79
Annif	Ensemble	C3	68,88	43,03	52,97
LLM	Pipeline híbrido	C3	34,73	21,64	26,67

Fonte: elaboração própria com base nos resultados experimentais

Os resultados indicam que os melhores desempenhos foram obtidos pelo SISA e pelo Annif–MLLM, que alcançaram resultados praticamente equivalentes e superiores às demais abordagens.

Em contraste, a abordagem baseada em LLM híbrida apresentou desempenho consideravelmente inferior, seus valores de Medida-F permaneceram significativamente abaixo daqueles observados para SISA e Annif. No entanto, essa diferença não deve ser interpretada de forma simplista como fracasso da abordagem generativa, mas como indício de que sua contribuição mais relevante pode não estar na substituição direta dos sistemas tradicionais, e sim em funções complementares dentro de arquiteturas híbridas. Além disso, os resultados

podem estar relacionados a limitações na configuração ou no ajuste do sistema, indicando a necessidade de novas investigações.

Tabela 2 – Documentos e f-measure organizados por intervalos

Sistema		f-m 0%	f-m 1-15%	f-m 16-30%	f-m 31-49%	f-m 50-74%	f-m ≥ 75%	Documentos
Annif	C1	0	1	19	194	272	14	500
	C2	0	0	2	68	386	44	500
	C3	2	0	0	61	378	59	500
SISA	C1	0	2	21	186	272	19	500
	C2	0	0	4	33	387	76	500
	C3	2	0	0	29	377	92	500
Llama 3	C1	12	129	256	91	12	0	500
	C2	14	117	221	126	22	0	500
	C3	17	100	208	148	27	0	500

Fonte: elaboração própria com base nos resultados experimentais

A Tabela 2 apresenta a distribuição das três subcoleções em seis intervalos percentuais, permitindo comparar a variação dos valores conforme a subcoleção e o sistema.

No Annif, a distribuição concentra-se nos intervalos mais altos, sobretudo entre 50–74%, com aumento desse intervalo em C2 e C3. O intervalo ≥75% também apresenta valores expressivos, indicando predominância de resultados altos.

SISA apresenta um comportamento ainda mais concentrado em valores elevados. Em C1, o intervalo 50–74% já é predominante e o intervalo ≥75% aumenta progressivamente em C2 e C3. Ao mesmo tempo, os valores baixos praticamente desaparecem nessas subcoleções, com uma concentração ainda mais acentuada nos valores altos do que em Annif.

Em contraste, a arquitetura baseada em LLM apresenta distribuição enviesada para valores baixos. Em C1, há 12 documentos em 0% e 129 em 1–15%, concentrando mais de 80% nos três primeiros intervalos. Em C2 e C3 há alguma migração para valores médios, mas os intervalos altos (≥50%) não aparecem, diferentemente de SISA e Annif.

Em síntese, o SISA apresenta o melhor desempenho, com maior concentração em valores altos (≥75%). O Annif ocupa posição intermédia, predominando em 50–74%. Já o LLM apresenta pior desempenho, com predominância de valores baixos e ausência de resultados altos.

4.2 INTERPRETAÇÃO COMPARATIVA DOS PARADIGMAS

Embora esta etapa preliminar esteja centrada na abordagem com LLMs, os resultados já permitem uma leitura comparativa mais ampla entre os três paradigmas centrais da pesquisa: SISA, Annif e LLM híbrida.

Tabela 3 – Leitura comparativa dos paradigmas de indexação automática analisados

Paradigma	Natureza da abordagem	Forças principais	Limitações principais	Papel mais promissor
SISA	Heurística / baseada em regras	Estabilidade terminológica; aderência documental; previsibilidade	Menor flexibilidade semântica; dependência de ocorrência textual explícita	Baseline documental e controle terminológico
Annif	Supervisionada / aprendizado de máquina	Capacidade de aprender padrões de indexação; escalabilidade; robustez em grandes conjuntos	Dependência de dados de treinamento; restrição a padrões já observados	Baseline estatística contemporânea
LLM híbrida	Generativa com recuperação lexical e semântica	Maior plasticidade semântica; compreensão contextual; apoio à seleção temática	Fragilidade terminológica; risco de sobregeração; variação por domínio	Apoio semântico, reranqueamento e indexação assistida

Fonte: elaboração própria com base nos resultados experimentais

4.3 O PAPEL DOS LLMS NA INDEXAÇÃO

Os resultados obtidos reforçam uma interpretação metodologicamente importante: no corpus de Agricultura analisado, o papel mais promissor da LLM não se mostrou o de indexadora final autossuficiente, mas o de componente de apoio semântico em arquiteturas híbridas. Essa interpretação converge com Liu *et al*. (2025), que combinam modelos baseados em embeddings com LLMs para controlar a geração de assuntos, e com Luo e Liu (2025), cuja proposta associa LLM, recuperação em vocabulário controlado e reranqueamento para reduzir alucinações terminológicas.

Essa leitura também é coerente com resultados institucionais recentes. Suominen, Inkinen e Lehtinen (2025) mostram que sistemas que mantêm algoritmos tradicionais de classificação e utilizam LLMs em funções específicas, como geração de dados e

reranqueamento, podem permanecer altamente competitivos. Dermentzi *et al*. (2025), por sua vez, concluem que, apesar do potencial das abordagens baseadas em LLM, os sistemas avaliados ainda são mais adequados a fluxos semiautomáticos com supervisão humana. Nos dados deste estudo, os níveis modestos de revocação e Medida-F da arquitetura LLM corroboram a distinção entre compreender semanticamente o documento e selecionar corretamente descritores autorizados.

A inferioridade quantitativa da abordagem baseada em LLM neste experimento não invalida seu interesse, mas delimita as condições em que sua contribuição se torna mais plausível. Włodarczyk (2025), em um corpus de artigos poloneses de Biblioteconomia e Ciência da Informação, observou desempenho semelhante entre GPT-4o e um serviço baseado em Annif, enquanto os resultados aqui obtidos foram claramente inferiores aos de SISA e Annif. A divergência entre os estudos sugere sensibilidade ao domínio, ao vocabulário, à composição do corpus e ao desenho do pipeline, tornando inadequada qualquer conclusão geral sobre superioridade ou inferioridade dos LLMs a partir de um único cenário experimental.

Em tarefas de indexação sob vocabulário controlado, portanto, a questão central não é apenas reconhecer sobre o que trata um documento, mas converter essa compreensão em termos autorizados, estáveis e úteis à representação documental. Os resultados de Luo e Liu (2025), Liu *et al*. (2025) e Suominen, Inkinen e Lehtinen (2025) indicam caminhos semelhantes ao adotado neste experimento: restringir ou recuperar previamente candidatos do vocabulário e reservar à LLM funções de seleção, refinamento ou reranqueamento. Nesse sentido, os achados do corpus agrícola analisado favorecem a investigação de arquiteturas híbridas, e não a substituição direta dos mecanismos especializados por geração livre.

5 CONSIDERAÇÕES FINAIS

Este trabalho comparou três abordagens de indexação automática sob vocabulário controlado - heurística, supervisionada e baseada em LLM - em um corpus específico de artigos científicos em inglês do domínio da Agricultura, utilizando descritores do NAL Agricultural Thesaurus e padrões de referência derivados da AGRICOLA.

No contexto experimental efetivamente investigado, SISA e Annif apresentaram desempenho superior à arquitetura baseada em LLM em precisão, revocação e Medida-F. Esse resultado não deve ser extrapolado para outros domínios ou configurações, pois reflete as

características do corpus, do vocabulário controlado e dos parâmetros utilizados no experimento.

Do ponto de vista da Organização do Conhecimento, os resultados reforçam, para este corpus, que a indexação automática não se limita à identificação temática do documento, mas envolve sua tradução para uma linguagem documental controlada. Como observam Poley *et al*. (2025), a automação em ambientes institucionais depende da produção de metadados temáticos suficientemente consistentes para integração aos fluxos de tratamento e recuperação da informação.

Os achados permitem destacar três pontos restritos ao cenário analisado: a arquitetura com Llama 3 mostrou viabilidade técnica para integrar recuperação lexical, recuperação semântica e reranqueamento; a imposição de candidatos autorizados evitou que a saída final utilizasse termos externos ao vocabulário; e, apesar desse controle, o desempenho quantitativo permaneceu inferior ao dos sistemas especializados, indicando que a ancoragem terminológica, por si só, não resolve a tarefa de seleção documental de descritores.

Entre as limitações do estudo estão a concentração em um único domínio, o uso de metadados em inglês provenientes de oito periódicos, a dependência de um único vocabulário controlado e de uma configuração específica do Llama 3, além da avaliação predominantemente quantitativa. Essas condições impedem generalizar os resultados para outras áreas do conhecimento, outros tipos documentais ou outros modelos de linguagem.

Em síntese, os resultados do corpus agrícola indicam que, nas condições testadas, abordagens tradicionais e supervisionadas permanecem mais eficazes para a atribuição de descritores do NAL, enquanto a principal contribuição potencial dos LLMs está em funções complementares de recuperação semântica, seleção e reranqueamento dentro de arquiteturas híbridas.

Como trabalhos futuros, propõe-se reproduzir o experimento em outros domínios e vocabulários controlados, incorporar documentos em texto integral, comparar diferentes modelos e estratégias de embeddings, avaliar alternativas de fusão e reranqueamento e incluir avaliação qualitativa por profissionais da indexação. Essa ampliação permitirá verificar se os padrões observados no domínio agrícola se mantêm ou se resultam de características específicas da coleção e do vocabulário empregados.

REFERÊNCIAS

ASULA, Marit; MAKKE, Jane; FREIENTHAL, Linda; KUULMETS, Hele-Andra; SIREL, Raul. Kratt: Developing an automatic subject indexing tool for the National Library of Estonia. **Cataloging & Classification Quarterly**, United States, v. 59, n. 8, p. 775-793, 2021. Disponível em: https://www.tandfonline.com/doi/full/10.1080/01639374.2021.1998283. Acesso em: 28 abr. 2026.

BOUZID, Sara; PIRON, Loïs. Leveraging generative AI in short document indexing. **Electronics**, Basel, v. 13, n. 17, art. 3563, 2024. Disponível em: https://www.mdpi.com/2079-9292/13/17/3563. Acesso em: 28 abr. 2026.

CHUN, Yongchan; KIM, Minhyuk; KIM, Dongjun; PARK, Chanjun; LIM, Heuiseok. Enhancing automatic term extraction with large language models via syntactic retrieval. *In*: FINDINGS OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS: ACL 2025, 2025, Vienna. **Anais** […]. Vienna: Association for Computational Linguistics, 2025. p. 9916-9926. Disponível em: https://aclanthology.org/2025.findings-acl.516/. Acesso em: 28 abr. 2026.

DERMENTZI, Maria; BRYANT, Mike; ROVIGO, Fabio; GARCÍA-GONZÁLEZ, Herminio. **Multilingual automated subject indexing**: a comparative study of LLMs vs alternative approaches in the context of the EHRI project. 2025. Working paper / preprint. Disponível em: https://hal.science/hal-05142136. Acesso em: 28 abr. 2026.

GIL-LEIVA, Isidoro. **Resources for the Evaluation of Automatic Indexing**: Corpora and Gold Standard Indexes in the Agricultural, Educational, and Medical Domains. [*S. l.*]: Zenodo, 2025. 1 conjunto de dados. Disponível em: https://doi.org/10.5281/zenodo.16440687. Acesso em: 6 set. 2026.

GIL-LEIVA, Isidoro. **Manual de indización**: teoría y práctica. Gijón: Trea, 2008.

GIL-LEIVA, Isidoro; DÍAZ ORTUÑO, Pedro; CORRÊA, Renato Fernandes. Automatic indexing of scientific articles on Library and Information Science with SISA, KEA and MAUI. **Revista Española de Documentación Científica**, Madrid, v. 45, n. 4, e338, 2022. Disponível em: https://redc.revistas.csic.es/index.php/redc/article/view/1371. Acesso em: 28 abr. 2026.

LIU, Jinyu; SONG, Xiaoying; ZHANG, Diana; THOMALE, Jason; HE, Daqing; HONG, Lingzi. **A hybrid framework for subject analysis**: integrating embedding-based regression models with large language models. 2025. Preprint. arXiv:2507.22913. Disponível em: https://arxiv.org/abs/2507.22913. Acesso em: 28 abr. 2026.

LUO, Pengcheng; LIU, Zhifeng. Automatic subject indexing with large language model fine-tuning and controlled vocabulary retrieval. **Aslib Journal of Information Management**, United Kingdom, 2025. Disponível em: https://www.emerald.com/ajim/article/doi/10.1108/AJIM-05-2025-0253/1307332/Automatic-subject-indexing-with-large-language. Acesso em: 28 abr. 2026.

POLEY, Christoph; UHLMANN, Sandro; BUSSE, Frank; JACOBS, Jan-Helge; KÄHLER, Maximilian; NAGELSCHMIDT, Matthias; SCHUMACHER, Markus. Automatic subject cataloguing at the German National Library. **Liber Quarterly**, Germany, v. 35, p. 1-29, 2025. Disponível em: https://liberquarterly.eu/article/view/19422. Acesso em: 28 abr. 2026.

RED DE BIBLIOTECAS UNIVERSITARIAS Y CIENTÍFICAS ESPAÑOLAS (REBIUN). **Anuario estadístico REBIUN 2022**. Madrid: REBIUN, 2023. Disponível em: https://hdl.handle.net/20.500.11967/1324. Acesso em: 6 set. 2026.

SUOMINEN, Osma. Annif: DIY automated subject indexing using multiple algorithms. **Liber Quarterly**, Germany, v. 29, n. 1, p. 1-25, 2019. Disponível em: https://liberquarterly.eu/article/view/10732. Acesso em: 28 abr. 2026.

SUOMINEN, Osma; INKINEN, Juho; LEHTINEN, Mona. Annif and Finto AI: developing and implementing automated subject indexing. **JLIS.It**, Firenze, v. 13, n. 1, p. 265-282, 2022. Disponível em: https://doi.org/10.4403/jlis.it-12740. Acesso em: 28 abr. 2026.

SUOMINEN, Osma; INKINEN, Juho; LEHTINEN, Mona. Annif at SemEval-2025 Task 5: Traditional XMTC augmented by LLMs. *In*: INTERNATIONAL WORKSHOP ON SEMANTIC EVALUATION (SemEval-2025). 19., 2025. **Anais** […]. Vienna: Association for Computational Linguistics, 2025. p. 2424-2431. Disponível em: https://aclanthology.org/2025.semeval-1.315.pdf. Acesso em: 28 abr. 2026.

UNITED STATES DEPARTMENT OF AGRICULTURE. National Agricultural Library. **NAL Agricultural Thesaurus and Glossary**. Washington, DC: USDA National Agricultural Library, 2017. Disponível em: https://lod.nal.usda.gov/nalt. Acesso em: 6 set. 2026.

WORLD BANK. **Scientific and technical journal articles**. World Bank Open Data, 2021. Disponível em: https://data.worldbank.org/indicator/IP.JRN.ARTC.SC. Acesso em: 6 set. 2026.

WŁODARCZYK, Bartłomiej. Automatic subject descriptions of Polish library and information science articles: a comparison of DESCRIPTOR e-service and GPT-4o. **Cataloging & Classification Quarterly**, [*s. l.*], v. 63, n. 6-7, p. 408-435, 2025. Disponível em: https://www.tandfonline.com/doi/full/10.1080/01639374.2025.2525841. Acesso em: 28 abr. 2026.