Revisão Sistemática sobre Expansão Automática de Léxicos Computacionais Assistida por LLMs

Luiz Felipe de Almeida e Silva, Graduando em sistemas de informação, UEG/CET, luizfelipedn57@aluno.ueg.br
Márcio Giovane Cunha Fernandes, Mestre, UEG/CET, marcio.giovane@ueg.br

Resumo: Este trabalho investiga os processos de formação automatizada de léxicos computacionais para Análise de Sentimentos (AS), no contexto do Processamento de Linguagem Natural (PLN). A pesquisa parte do problema da escassez de metodologias sistemáticas e culturalmente adequadas para construção de léxicos afetivos, frequentemente traduzidos de outros idiomas e com baixa acurácia semântica. Assume-se que os Modelos de Linguagem de Grande Escala (LLMs) podem facilitar a geração e curadoria desses léxicos, proporcionando escalabilidade e adaptação ao contexto linguístico e cultural.. O objetivo é analisar os processos descritos na literatura, avaliando sua aplicabilidade, replicabilidade em português brasileiro e possibilidades de automatização. A partir de uma Revisão Sistemática da Literatura (RSL), observa-se um campo ainda incipiente, mas em expansão, com resultados preliminares que apontam para uso ainda incipiente de LLMs na formação de léxicos computacionais precisos e contextualizados.

Palavras–chave: Léxico; Análise de Sentimentos; Processamento de Linguagem Natural; Modelos de Linguagem de Grande Escala.

INTRODUÇÃO

A Análise de Sentimentos (AS) tem se consolidado como uma das aplicações mais relevantes do Processamento de Linguagem Natural (PLN), especialmente diante do rápido crescimento de dados textuais oriundos de redes sociais, fóruns e avaliações de produtos (MARTINS, 2020). Ao buscar identificar automaticamente a polaridade e as emoções expressas em textos, a AS depende fortemente de léxicos computacionais — conjuntos de palavras associadas a valores afetivos. No entanto, a literatura revela uma significativa lacuna quanto à disponibilidade de léxicos robustos e sensíveis às especificidades linguísticas e culturais de línguas diferentes da língua inglesa. Grande parte dos léxicos utilizados consistem em traduções adaptadas de léxicos originalmente desenvolvidos para o inglês, comprometendo sua eficácia em outros contextos linguísticos.

Neste cenário, o presente trabalho tem como tema a formação automatizada de léxicos computacionais e como objeto específico a avaliação de processos que empregam Modelos de Linguagem de Grande Escala (LLMs) para a geração ou curadoria desses léxicos. O recorte teórico-metodológico envolve uma Revisão Sistemática da Literatura (RSL) focada em técnicas assistidas por LLMs. Trabalhos como os de Sousa (2023) e Rodrigues (2024) fornecem importantes referências ao explorarem metodologias sensíveis ao contexto brasileiro, mas que não exploram ferramentas de LLM, que vem se democratizando nos últimos anos.

A justificativa para a pesquisa reside na urgência por métodos replicáveis, escaláveis e linguística e culturalmente adequados para a construção de léxicos, sendo o objetivo geral identificar a tendência da produção científica internacional a respeito da expansão automática e assistida por LLMs de léxicos computacionais através de uma RSL.

PROCEDIMENTOS DE TRABALHO

Este trabalho adotou uma uma RSL para mapear, analisar e sintetizar processos automatizados de construção de léxicos computacionais assistidos por LLMs, aplicáveis ao contexto da AS em língua portuguesa. A metodologia da RSL seguiu o protocolo definido por Kitchenham (2013), assegurando rigor, transparência e reprodutibilidade. Inicialmente, definiu-se o conjunto de questões norteadoras da pesquisa: (i) “existe algum processo elaborado para construção

automática de léxico computacional usando LLM?", (ii) "quais são os métodos existentes para criação automatizada de léxico computacional?", (iii) "quais métricas são usadas para avaliar a qualidade desses léxicos?", (iv) "quais etapas na criação de um léxico computacional podem ser automatizadas?" e (v) "quais etapas na criação de um léxico computacional podem ser auxiliadas pelo LLM?"

A coleta de dados foi realizada em abril de 2025, em três bases científicas: *Scopus, IEEE Xplore* e *ACM Digital Library*. Utilizou-se uma expressão de busca estruturada com operadores booleanos e termos-chave como "*lexicon*", "LLM", "*automatic*", "*lexicon generation*" e seus sinônimos. A busca resultou em um *corpus* — conjunto de textos sobre um determinado tema — inicial de 165 documentos. Foram então aplicados os critérios de exclusão, sendo (i) o documento é uma versão duplicada de outro, (ii) o documento não é um artigo ou documento de conferência, (iii) o documento está fora do escopo deste projeto, (iv) o documento foi publicado antes de 2020 (justificado pelo ano de lançamento do *Chat GPT*), (v) o texto completo do artigo não está disponível para acesso e (vi) o texto completo do documento não está escrito em inglês ou português.

Os restantes não excluídos deveriam atender a pelo menos um critério de inclusão dentre os seguintes: (i) o documento contribui para discussões teóricas ou técnicas sobre geração automatizada de léxico computacional no contexto da PLN, (ii) o documento discute a criação de léxico computacional, (iii) o documento apresenta uma ou mais técnicas para geração de léxicos computacionais e (iv) a pesquisa utiliza LLMs para geração ou expansão de léxico computacional. Após essa triagem, restaram 18 documentos que compõem o *corpus* final da revisão, que passou por análises quantitativas e qualitativas, que serão dispostos na próxima seção deste trabalho. Adicionalmente, realizamos uma análise quantitativa do *corpus* inicial da busca, que não passou por critérios, para comparação entre os resultados. Os critérios foram refinados durante a aplicação, conforme Kitchenham (2013), garantindo rigor e adaptabilidade.

RESULTADOS

Os dados, extraídos da busca sistemática inicial permitiram a caracterização quantitativa do *corpus* bibliográfico. Foram analisados indicadores de produtividade acadêmica, incluindo distribuição temporal dos documentos, identificando padrões de crescimento, núcleos de excelência científica, mediante mapeamento de autores e instituições com maior densidade de publicações, periódicos e eventos de maior relevância temática, análise de cocitação e acoplamento bibliográfico para detecção de redes de influência intelectual, palavras chave mais recorrentes e países com maior impacto na produção científica do tema.

As análises do *corpus* inicial de busca demonstraram um crescimento linear de frequência de ocorrências de termos como *Deep Learning*, *Embeddings* (técnica de representação numérica sobre um termo ou expressão linguística), *Sentiment Analysis* e *Computational Linguistics* a partir de 2020 e 2021, com destaque para os dois últimos, que apresentaram crescimento mais acentuado na frequência de uso. Adicionalmente, foi possível identificar os países mais influentes, destacando a China, Índia e Estados Unidos, enquanto não há ocorrências de produções brasileiras neste tema. Por fim, a análise do *corpus* inicial da busca também demonstrou um crescimento expressivo de publicações relacionadas ao tema, conforme demonstra a Figura 1, onde é possível observar a tendência em constante crescente a partir do ano de 2021.

No que diz respeito ao *corpus* selecionado, que contém os documentos aceitos pelos critérios de aceitação, identifica-se os Estados Unidos e Inglaterra com a maior quantidade de produções sobre o tema delimitado, além de um enfoque maior em termos como "lexicon" em comparação ao *corpus* inicial, conforme Figura 2.

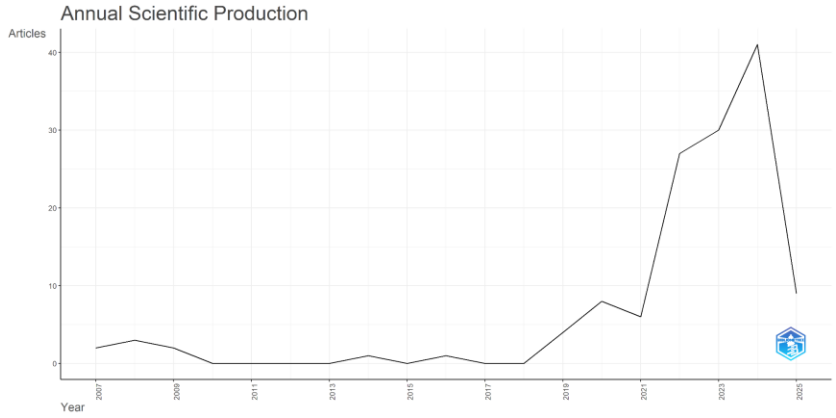

Figura 1: Tendência da produção científica anual de 2007 a 2025, extraída do *corpus* inicial da busca

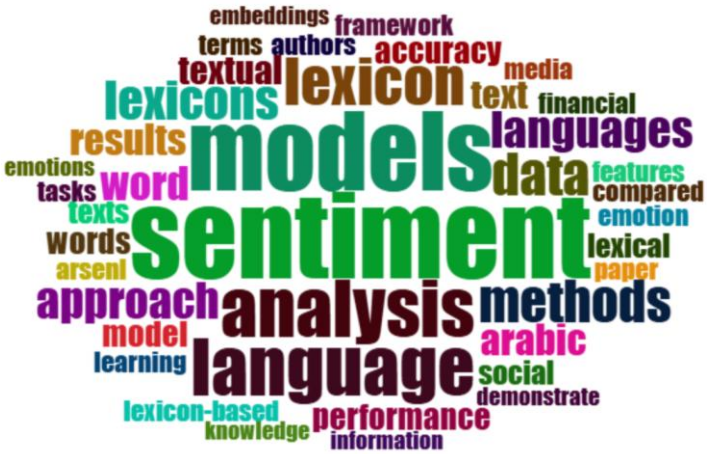

Figura 2: Nuvem de palavras extraídas dos resumos dos documentos do *corpus* final.

No que diz respeito a análise qualitativa, foi possível identificar uma preocupação internacional em manter a orientação binária (positiva ou negativa) e sentimental (raiva, medo, otimismo, satisfação) em textos traduzidos automaticamente e para isso, diversas técnicas estão sendo aplicadas com essa finalidade, unindo utilização de léxicos computacionais com LLMs, sendo esta uma temática fora do escopo deste trabalho, que busca observar processos de expansão automatizada de léxicos (o que justificou os critérios de inclusão e exclusão elaborados). Sobre os trabalhos selecionados, destacamos trabalhos como os de Rizinski *et a*l. (2023) e Du *et a*l. (2023) que desenvolveram seus próprios processos utilizando técnicas e abordagens diferentes para expansão ou criação de léxicos computacionais para AS para mercado financeiro. Também destaca-se o trabalho de Nakiwjit, Samir e Purver (2023), que fez uso de LLMs como *ChatGPT* para refinamento do léxico.

DISCUSSÃO

Este estudo partiu do pressuposto de que LLMs podem otimizar a construção e expansão de léxicos computacionais para AS, superando limitações como a escassez de recursos nativos e a inadequação cultural de léxicos traduzidos. Os resultados da RSL confirmaram parcialmente essa hipótese, revelando um campo em ascensão, mas ainda incipiente. A análise bibliométrica demonstrou um crescimento exponencial de publicações sobre o tema a partir de 2021, concentradas majoritariamente em países como China, Índia e Estados Unidos, enquanto a

produção brasileira permanece ausente no *corpus* analisado, demonstrando também a deficiência em línguas menos populares.

A análise qualitativa dos 18 artigos selecionados evidenciou que os LLMs — como *ChatGPT* e *Gemini* — são pouco empregados, concentrando-se, quando usados, principalmente em etapas de refinamento lexical. No entanto, persistem desafios críticos: (i) a maioria dos métodos é desenvolvida para línguas com recursos abundantes (como o inglês), sem adaptação sistemática para variações linguísticas de outras línguas; (ii) faltam métricas padronizadas para avaliar a qualidade cultural e semântica dos léxicos gerados; e (iii) há pouca exploração de LLMs em domínios específicos. A metodologia adotada — baseada nos critérios de Kitchenham (2013) — mostrou-se adequada para mapear tendências e lacunas, mas revelou limitações práticas para contextos nacionais específicos, como no caso do Brasil, tal como outros países falantes de língua portuguesa.

Como desdobramentos, sugere-se: (i) a realização de experimentos dos processos propostos em corpora em português, testando sua eficácia na captura de nuances culturais, conforme sugere Rodrigues (2024); (ii) a identificação de métricas para avaliação de léxicos gerados automaticamente; e (iii) a investigação de métodos que integrem LLMs. Este trabalho, ao sistematizar o estado da arte, oferece um ponto de partida para futuras pesquisas que busquem democratizar o acesso a recursos de PLN para línguas menos representadas.

CONCLUSÕES

Conclui-se que a formação automatizada de léxicos computacionais assistida por LLMs é um campo promissor, mas ainda subexplorado, especialmente em contextos linguísticos diversos como o português brasileiro. A sistematização realizada permite compreender os caminhos já trilhados e evidencia a necessidade de metodologias adaptáveis e sensíveis ao contexto. Este trabalho contribui ao oferecer uma base teórico-metodológica sólida para futuras investigações, destacando lacunas e propondo direções estratégicas para pesquisas aplicadas em AS, com potencial de ampliar a inclusão linguística e cultural nos avanços do PLN.

REFERÊNCIAS

DE SOUSA, Thiago Monteles; FERNANDES, Deborah S. A. Expansão automática de léxico para Análise de Sentimentos de Twitter no domínio do Mercado Financeiro Brasileiro. In: ESCOLA REGIONAL DE INFORMÁTICA DE GOIÁS (ERI-GO), 11., 2023, Goiânia. **Anais** [...]. Porto Alegre: Sociedade Brasileira de Computação, 2023.

DU, K., XING, F., MAO, R., *et al*. FinSenticNet: A Concept-Level Lexicon for Financial Sentiment Analysis. **IEEE Symposium Series on Computational Intelligence, SSCI 2023**. Institute of Electrical and Electronics Engineers Inc. Disponível em: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85178663623&doi=10.1109%2fSSCI52147.2023.10371970&partnerID=40&md5=1ede88a2a7848d0ba6de42bf60f180f9. Acesso em: 24 abr. 2025.

KITCHENHAM, B.; BRERETON, P. A systematic review of systematic review process research in software engineering. **Information and Software Technology**, v. 55, n. 12, p. 2049–2075, 2013. Disponível em: https://www.sciencedirect.com/science/article/pii/S0950584913001560. Acesso em: 24 abr. 2025.

MARTINS, J. S. *et a*l. **Processamentos de Linguagem Natural**. Porto Alegre: SAGAH, 2020. E-book. p.251 266. Disponível em: https://app.minhabiblioteca.com.br/reader/books/9786556900575/.

NAKWIJIT, P., SAMIR, M., PURVER, M. Lexicools at SemEval-2023 Task 10: Sexism Lexicon Construction via XAI. **17th International Workshop on Semantic Evaluation, SemEval 2023 - Proceedings of the Workshop**, Association for Computational Linguistics. Disponível em: https://www.scopus.com/inward/record.uri?eid=2-s2.0-

85175399351&doi=10.18653%2fv1%2f2023.semeval-1.4&partnerID=40&md5=ef37c29fc0ad82a89b8177061eebf805. Acesso em: 24 abr. 2025

RIZINSKI, M. *et al*. Sentiment Analysis in Finance: From Transformers Back to eXplainable Lexicons (XLex). **IEEE access**, v. 12, p. 7170–7198, 2024. Disponível em: https://www.scopus.com/record/display.uri?eid=2-s2.0-85182367770&doi=10.1109%2fACCESS.2024.3349970&origin=inward&txGid=1c37021206b5740211744e8a9c80656b. Acesso em 24 abr. 2025

RODRIGUES, Leidiane; FERNANDES, Deborah; DO LAGO, Marilúcia Pereira; FERNANDES, Márcio; SOARES, Fabrizzio; SILVA, Kairo. Visão sobre técnicas computacionais na detecção de depressão em texto. **Journal of Health Informatics**, Brasil, v. 16, n. Especial, 2024. DOI: 10.59681/2175-4411.v16.iEspecial.2024.1363. Disponível em: https://jhi.sbis.org.br/index.php/jhi-sbis/article/view/1363. Acesso em: 24 abr. 2025.