



Identificação de Padrões Linguísticos na Escrita Científica por meio de Técnicas de Processamento de Linguagem Natural

Gabriel Evaristo Santana da Silva¹, Ionildo José Sanches¹

¹ Programa de Pós-graduação em Ciência da Computação - Universidade Tecnológica Federal do Paraná (UTFPR), Ponta Grossa, Brasil
(gabrielevaristo@alunos.utfpr.edu.br)

Resumo: A escrita acadêmica é essencial à pesquisa e à educação, pois organiza formas de expressar ideias e comunicar ciência. Este trabalho investiga padrões da escrita científica com técnicas de Processamento de Linguagem Natural aplicadas a 800 artigos de Computação (2021–2024). Foram desenvolvidos *pipelines* de extração, pré-processamento e análise textual para identificar padrões temáticos e estruturais. Os resultados indicam regularidades na escrita acadêmica.

Palavras-chave: Escrita Científica; Comunicação Científica; Processamento de Linguagem Natural.

INTRODUÇÃO

Amplamente utilizada por pesquisadores e educadores em trabalhos acadêmicos para apresentar argumentos baseados em dados e raciocínios lógicos, a escrita acadêmica é um aspecto fundamental da pesquisa e da educação, estruturando métodos para expressar ideias e comunicar a ciência. Dessa forma, auxilia leitores a compreender a pesquisa de forma aprofundada e permite que autores analisem conceitos e produzam resultados, teorias e conclusões bem fundamentadas. Por esses motivos, a escrita acadêmica científica é utilizada em diversas áreas do conhecimento (Rodrigues, 2020; Younis *et al.*, 2023).

É fundamental considerar que a divulgação científica, a partir da escrita acadêmica, deve ser direcionada a públicos variados, frequentemente provenientes de diferentes áreas do conhecimento. Neste contexto, ferramentas de Inteligência Artificial (IA) podem apoiar escritores na estruturação, edição e refinamento de textos científicos. Abrangendo uma ampla variedade de funcionalidades, a IA pode realizar funções que eram associadas apenas à mente humana (Younis *et al.*, 2023; Hutson e Plate, 2022).

Neste contexto, um dos campos da IA que mais se destaca é o Processamento de Linguagem Natural (PLN), que inclui a capacidade da IA de processar e compreender a comunicação escrita e oral, podendo identificar estruturas ocultas ou padrões em grandes conjuntos de dados baseados em texto (Caseli e Nunes, 2024). No contexto educacional, o PLN tem sido um tema de interesse com pesquisas nas áreas de correção e *feedback* automatizados e na análise de redações escritas por estudantes (Berdanier, McComb e Zhu, 2020).

Dessa forma, com algoritmos de PLN é possível desenvolver modelos computacionais para tarefas que dependem de informações da linguagem natural, como por exemplo tradução de texto e busca de informações em documentos. Assim, os modelos desenvolvidos são utilizados para o estudo da linguagem dos seres humanos, propiciando a criação de ferramentas que permitam a compreensão, a análise e a geração de linguagem natural (Pinto, 2025).

O PLN relaciona-se a diversas áreas de estudos, sendo uma área interdisciplinar entre linguística humana e computacional, estatística, inteligência artificial e aprendizado de máquina, mineração de dados, reconhecimento e síntese da fala humana, entre outros. Além disso, pode realizar o processamento da linguagem humana de forma automática ou semi-automática através da interação homem-máquina, gerando informações, sumarização de textos, análise de sentimentos, reconhecimento de fala, mineração de dados, aprendizado profundo e *chatbots* (Lee, 2023).

Diante do exposto, observa-se que a comunicação científica, a escrita acadêmica e as ferramentas de IA, em especial o PLN, têm se mostrado elementos centrais para o aprimoramento da produção e disseminação do conhecimento. Considerando o crescente interesse em aplicar algoritmos de PLN para analisar e identificar padrões na escrita científica, o objetivo deste trabalho é investigar e identificar padrões recorrentes na escrita científica, com o auxílio de técnicas de PLN, com o intuito de desenvolver subsídios e estratégias que apoiem autores na produção de textos acadêmicos mais claros, coerentes e adequados às normas da comunicação científica.



Para alcançar esse objetivo, propõe-se analisar produções acadêmicas na área da computação com o intuito de identificar estruturas, vocabulários e padrões linguísticos comuns, bem como aplicar técnicas de PLN para extrair, classificar e interpretar elementos recorrentes em textos científicos, como conectores lógicos e estruturas argumentativas.

Além disso, busca-se avaliar a eficácia de ferramentas baseadas em IA na detecção e recomendação de padrões de escrita científica, propor diretrizes ou modelos de apoio à escrita acadêmica com base nos padrões identificados, visando auxiliar estudantes, pesquisadores e educadores, e, por fim, explorar as potencialidades pedagógicas do uso de ferramentas de IA e PLN no ensino da escrita científica, especialmente no contexto da comunicação científica como disciplina formativa.

MATERIAL E MÉTODOS

O presente trabalho caracteriza-se pelo desenvolvimento de um modelo de PLN para análise de artigos científicos. O objetivo é identificar padrões de escrita científica, contribuindo para auxiliar estudantes e pesquisadores na elaboração e comunicação de trabalhos acadêmicos. A pesquisa possui caráter descritivo e abordagem qualitativa, sendo estruturada em três fases principais: pesquisa descritiva, experimental e analítica.

Na fase descritiva, realizou-se o mapeamento e a interpretação de estudos já existentes, buscando compreender padrões de escrita científica e metodologias adotadas na literatura.

Em seguida, na fase experimental, foi realizada a coleta de dados na Plataforma Sucupira, considerando periódicos da área de Computação qualificados no estrato A1 no Qualis CAPES (2021–2024). A partir desse recorte, foram selecionados 800 artigos científicos publicados entre 2021 e 2024, garantindo a utilização de trabalhos recentes e de alto impacto. A distribuição dos artigos evidencia predominância de publicações no ano de 2024 (78%), reforçando a atualidade do *corpus* analisado.

Na fase analítica, para viabilizar a análise computacional, foi desenvolvido um primeiro *pipeline* para extração textual dos artigos em formato PDF (*Portable Document Format*, em português Formato de Documento Portátil). Utilizando *Python* no ambiente *Google Colab*, os arquivos foram armazenados e processados, realizando a leitura página a página e a extração do conteúdo textual. O processo incluiu tratamento de exceções e execução paralela, otimizando o desempenho. Os dados extraídos foram organizados em um *dataset* estruturado contendo identificador, texto completo e *status* de processamento, sendo posteriormente

exportados em formato CSV (*Comma Separated Values*, em português Valores Separados por Vírgula).

O segundo *pipeline* foi dividido em três etapas: carregamento dos dados, pré-processamento textual e análise textual. Na etapa de carregamento, os dados foram importados e submetidos a verificações exploratórias, como a visualização das primeiras linhas e a conferência da dimensão do *dataset*, a fim de garantir a correta importação e integridade das informações.

Em seguida, foi realizada a etapa de pré-processamento textual, com conversão dos textos para formato uniforme, correção de caracteres corrompidos oriundos da extração dos PDFs, remoção de quebras de linha, caracteres especiais, pontuações indesejadas e espaços excedentes, além da padronização do texto em letras minúsculas. Esse processo permitiu preparar os dados para as análises posteriores.

Na etapa de análise textual, foram inicialmente identificadas as palavras mais frequentes do *corpus*. Para isso, os textos pré-processados foram reunidos e tokenizados, processo no qual o conteúdo textual é dividido em unidades menores denominadas *tokens*, geralmente correspondentes a palavras individuais. Para essa etapa, foi utilizada a biblioteca *Natural Language Toolkit* (NLTK), uma das principais bibliotecas de processamento de linguagem natural em Python (Bird, Klein e Loper, 2009). Em seguida, os textos foram filtrados por meio da remoção de *stopwords* em inglês, números, símbolos e termos muito curtos. Em seguida, foi calculada a frequência dos termos, permitindo identificar as 30 palavras mais recorrentes. Também foi realizada a extração automática de *keywords*, considerando variações como *keywords*, *key words*, *key words and phrases* e *index terms*. As palavras-chave foram separadas por vírgulas e ponto e vírgula, normalizadas em letras minúsculas e tratadas para unificar termos equivalentes, como *deep learning* e *deeplearning*, evitando duplicidades semânticas. Após esse processo, foram identificadas as 20 *keywords* mais frequentes.

Complementarmente, o *pipeline* incluiu a extração de *bigrams*, isto é, pares de palavras consecutivas, com o objetivo de identificar padrões de coocorrência textual mais complexos (O'Shaughnessy, 2026). Essa etapa envolveu nova filtragem dos textos, com remoção de URLs, números, *stopwords*, termos ruidosos associados a metadados editoriais, elementos de PDF, métricas de publicação, instituições e referências a figuras. Também foram aplicados critérios de tamanho das palavras e validação dos termos para reduzir ruídos e inconsistências. Os *bigrams* resultantes foram contabilizados, permitindo identificar as 15 combinações mais recorrentes no *corpus*. Por fim, foi realizada a detecção automática de seções estruturais

RESULTADOS E DISCUSSÃO

A análise de frequência de palavras revelou a predominância de termos associados à inteligência artificial, aprendizado de máquina e sistemas computacionais. Como mostra a Figura 1, destacam-se palavras como *model*, *data*, *learning*, *network* e *algorithm*, indicando forte concentração temática nessas áreas. Além disso, a recorrência de termos como *method*, *approach*, *proposed* e *evaluation* evidencia a presença de um padrão discursivo consolidado, caracterizado pela proposição de soluções, descrição metodológica e validação de resultados. Esse comportamento reforça a hipótese de que a escrita científica segue estruturas linguísticas relativamente estáveis, especialmente em periódicos de alto impacto.

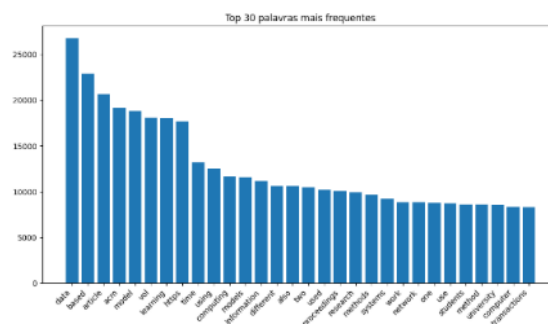


Figura 1. As 30 palavras com maior frequência no *corpus*.

keywords	Frequência
deep learning	35
machine learning	25
federated learning	20
privacy	18
blockchain	14
covid-19	13
recommender systems	12
security	11
software testing	11
neural networks	10
internet of things	10
largelanguage models	9
computer science education	9
edge computing	8
representation learning	7
computational thinking	7
computing education	6
reinforcement learning	5
graph neural networks	5
contrastive learning	5

Figura 2. As 20 *keywords* com maior frequência no *corpus*.

A análise de *bigrams* permitiu aprofundar a compreensão das relações semânticas entre os termos, evidenciando combinações frequentes que reforçam os principais eixos temáticos do *corpus*. Como mostra a Figura 3, destacam-se expressões como *neural networks*, *machine learning*, *computer vision* e *software engineering*, além de combinações como *real time* e *real world*, associadas a aplicações práticas. Esses resultados reforçam a coerência entre as análises de palavras e de palavras-chave, mostrando não apenas os temas predominantes, mas também a forma como os conceitos se articulam nos textos científicos.

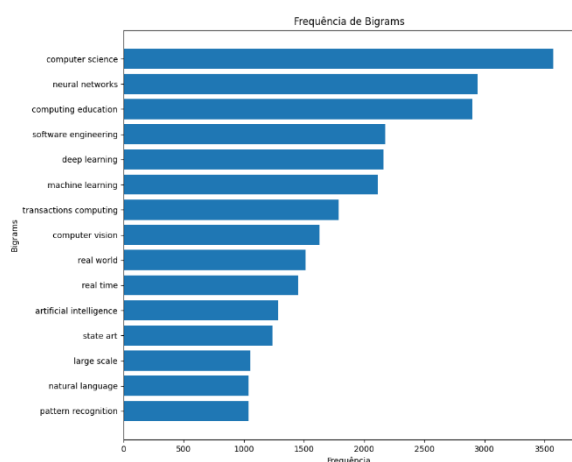


Figura 3. Os 15 *bigrams* com maior frequência no *corpus*.

No que se refere à estrutura dos artigos, a análise das seções indicou forte aderência ao modelo IMRaD, evidenciada pela alta frequência de termos associados às principais partes da organização científica. A Tabela 1 apresenta a distribuição das seções identificadas no *corpus*, indicando quantas vezes cada título ou variação de seção apareceu nos artigos analisados. Observa-se maior ocorrência de *methods* e *results*, com 19.851 e 10.906 registros, respectivamente, o que sugere forte ênfase na descrição metodológica e na apresentação de resultados. Também foram identificadas seções como *introduction*, *discussion* e *conclusion*, reforçando a presença de uma estrutura textual convencional. Por outro lado, a ocorrência de termos como *background*, *related work* e *related reviews* indica certa flexibilidade na nomenclatura e na organização das seções, especialmente em artigos da área de Computação.

Tabela 1. Frequência das seções identificadas no *corpus* de artigos científicos analisados.

Seção	Frequência
<i>methods</i>	19851
<i>results</i>	10906
<i>introduction</i>	1275
<i>background</i>	1205
<i>discussion</i>	1045
<i>conclusion</i>	989
<i>abstract</i>	693
<i>related work</i>	525
<i>related reviews</i>	10

De forma geral, os resultados evidenciam uma convergência entre padrões temáticos e estruturais na escrita científica em Computação, com destaque para a predominância de pesquisas em inteligência artificial e a adoção de estruturas discursivas consolidadas. A integração das análises de frequência de palavras, palavras-chave e *bigrams* demonstrou-se eficaz para a identificação de padrões temáticos, contribuindo para uma compreensão mais ampla da escrita acadêmica na área. Esses achados podem subsidiar futuras investigações e aplicações voltadas à análise automatizada de textos científicos e ao apoio à produção acadêmica.

Entretanto, este trabalho encontra-se em andamento, e os resultados apresentados correspondem a uma etapa inicial da pesquisa. Como continuidade, pretende-se ampliar o conjunto de técnicas de PLN aplicadas, incorporando abordagens mais avançadas para aprofundar a análise dos padrões textuais. Entre as próximas etapas, destacam-se a aplicação de modelos para análise semântica mais refinada e a implementação de modelos baseados em *transformers* (como BERT) para captura de relações contextuais mais complexas.

CONCLUSÃO

Conclui-se que a aplicação de técnicas de PLN possibilitou a identificação de padrões temáticos e estruturais relevantes na escrita científica da área de Computação. As análises realizadas evidenciaram a predominância de temas associados à inteligência artificial, ao aprendizado de máquina e aos sistemas computacionais, além de indicarem regularidades discursivas e forte aderência ao modelo IMRaD. Esses resultados demonstram que abordagens baseadas em frequência de palavras, palavras-chave, *bigrams* e identificação de seções podem contribuir para compreender a organização textual e os padrões recorrentes da comunicação científica.

Como continuidade, pretende-se ampliar as análises por meio de técnicas mais avançadas de PLN, incluindo classificação textual, análise de similaridade, detecção de padrões argumentativos, avaliação de coesão e coerência, extração de relações semânticas e identificação de conectores discursivos. Espera-se que essas etapas possibilitem uma identificação mais precisa dos padrões da escrita acadêmica e subsidiem o desenvolvimento de diretrizes e ferramentas inteligentes capazes de apoiar estudantes, pesquisadores e educadores na produção de textos científicos mais claros, estruturados e adequados às práticas da comunicação acadêmica.



REFERÊNCIAS

- BERDANIER, Catherine; MCCOMB, Christopher; ZHU, Weiwei. Natural Language Processing for Theoretical Framework Selection in Engineering Education Research. *Frontiers in Education Conference (FIE)*, p. 1–7, 2020.
- BIRD, Steven; KLEIN, Ewan; LOPER, Edward. *Natural Language Processing with Python*. [S.l.]: O'Reilly Media, 2009.
- CASELI, H. M.; NUNES, M. G. V. (Ed.). *Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português*. 2. ed. BPLN, 2024. Disponível em: <<https://brasileiraspln.com/livro-pln/2a-edicao/>>. Acesso em: 17 jan. 2026.
- HUTSON, James; PLATE, Daniel. *Augmented Creativity: Leveraging Natural Language Processing for Creative Writing*. *Art and Design Review*, v. 10, n. 08, 2022.
- LEE, Raymond. *Natural Language Processing: A Textbook with Python Implementation*. [S.l.: s.n.], 2023.
- O'SHAUGHNESSY, Douglas. An overview of recent advances in natural language processing for information systems. *Applied Sciences*, v. 16, n. 2, 2026. ISSN 2076-3417. Disponível em: <<https://www.mdpi.com/2076-3417/16/2/1122>>. Acesso em: 20 nov. 2025.
- PINTO, Sara Catarina Silva. *Processamento de Linguagem Natural e Extração de Conhecimento*. Dissertação (Tese de Mestrado) — Universidade de Coimbra, Departamento de Engenharia Informática, Faculdade de Ciências e Tecnologias, jul. 2015. Disponível em: <<https://estudogeral.uc.pt/handle/10316/35676>>. Acesso em: 7 mar. 2026.
- RODRIGUES, Rosana Ferrareto Lourenço. A ciência é uma jornada: um projeto remodelado como programa de pesquisa linguística em comunicação científica com uso de data science. *Sinergia (IFSP)*, p. 60–81, 2020. Disponível em: <<https://ojs.ifsp.edu.br/index.php/sinergia/article/view/1111>>. Acesso em: 10 jul. 2025.
- YOUNIS, Hussain A. *et al.* A Systematic Literature Review on the Applications of Robots and Natural Language Processing in Education. *Electronics*, v. 12, n. 13, 2023.