AVALIAÇÃO COMPARATIVA DE MÉTODOS DE CLASSIFICAÇÃO DE ESPECIALIDADE MÉDICA EM CORPUS BIOMÉDICO COVID-19 EM PORTUGUÊS

Wellington Araujo Nogueira, André Kazuo Takahata[2]

[1] Centro de Engenharia, Modelagem e Ciências Sociais Aplicadas Universidade Federal do ABC (UFABC) - Santo André - SP – Brasil (wellington.araujo@aluno.ufabc.edu.br)
[2] Centro de Engenharia, Modelagem e Ciências Sociais Aplicadas Universidade Federal do ABC (UFABC) - Santo André - SP – Brasil

Resumo: A classificação automática de artigos biomédicos por especialidade médica constitui desafio relevante para a organização de repositórios científicos em língua portuguesa. Este trabalho compara seis métodos de classificação automática de artigos sobre COVID-19 em 21 especialidades médicas: um baseline por correspondência de palavras-chave e cinco estratégias de prompting com o modelo Claude Haiku 4.5 (zero-shot, one-shot e few-shot, com texto completo ou título). O corpus reúne 1.903 artigos do PubMed/MEDLINE, com gabarito de 178 artigos anotados manualmente. Todos os métodos baseados em grandes modelos de linguagem (LLM) superaram significativamente o baseline (F1 Macro: 0,106 → 0,567), um ganho de cerca de 5×. O few-shot com texto completo alcançou o melhor F1 Macro (0,567), mas com custo 15 vezes maior que o zero-shot, que apresentou a melhor relação eficiência-custo. O Cohen's Kappa indica concordância moderada dos métodos LLM com o gabarito (κ = 0,50–0,54). Conclui-se que o zero-shot oferece o melhor equilíbrio entre desempenho e custo computacional para este domínio.

Palavras-chave: Processamento de Linguagem Natural; Classificação Biomédica; Modelos de Linguagem de Grande Escala; COVID-19; Prompting; Corpus em Português.

INTRODUÇÃO

A pandemia de COVID-19 gerou enorme produção científica biomédica, inclusive em países de língua portuguesa, o que exige sistemas automáticos de categorização por especialidade médica. Iniciativas como CORD-19 e TREC-COVID atenderam à literatura em inglês, mas faltam soluções equivalentes para o português. No Brasil, Leite et al. [2020] criaram um corpus de COVID-19 a partir do PubMed e Assis et al. [2021] analisaram seu léxico biomédico, base direta deste estudo, embora a classificação automática por especialidade ainda não tenha sido avaliada com estratégias modernas de prompting com LLMs. Métodos por palavras-chave são simples, mas não capturam contexto semântico; abordagens como BERT e BioBERT exigem grandes conjuntos anotados, escassos em português. Os LLMs com aprendizado em contexto permitem classificação sem fine-tuning, via prompts zero-shot, one-shot ou few-shot. O trabalho investiga duas questões: se o prompting com LLMs supera a classificação por palavras-chave e qual estratégia oferece o melhor equilíbrio entre desempenho e custo. Para isso, avalia seis métodos (um baseline e cinco com Claude Haiku 4.5) sobre 1.903 artigos do PubMed/MEDLINE, com gabarito de 178 artigos em 21 especialidades, usando F1 Macro, Cohen's Kappa, MCC e intervalos de confiança bootstrap.

MATERIAL E MÉTODOS

2.1 Corpus e Gabarito

O corpus reúne 1.903 artigos sobre COVID-19 em revistas de língua portuguesa, extraídos do PubMed/MEDLINE via API E-utilities do NCBI, com query de termos MeSH e filtro de idioma, cobrindo do início da pandemia até abril de 2026. O pipeline expande a metodologia de Leite et al. [2020] e Assis et al. [2021], adicionando extração de texto completo via PubMed Central e download de PDFs em acesso aberto. De cada artigo extraíram-se título, resumo, palavras-chave, revista, autores, afiliações, DOI e data, submetidos a um pré-processamento de 8 etapas (remoção de boilerplate, URLs, datas, referências, stopwords, caracteres especiais e normalização de espaços). O gabarito tem 178 artigos anotados manualmente por especialistas em 21 especialidades, com distribuição desbalanceada (Epidemiologia e Cardiologia dominam, enquanto 8 especialidades têm 1 a 2 exemplos), o que impacta as métricas macro.

2.2 Métodos de Classificação

Foram avaliados seis métodos. O baseline (M0) usa um dicionário de 20 especialidades com 15 a 40 palavras-chave bilíngues por categoria e pontuação ponderada por campo (título e palavras-chave 3×, resumo 2×, corpo 1×). Os demais usam o Claude Haiku 4.5 via API Anthropic: (M1) zero-shot, apenas descrição da tarefa e texto; (M2) one-shot com texto completo; (M3) one-shot com título; (M4) few-shot com texto completo (até cinco exemplos por especialidade); (M5) few-shot com título. Em todos os métodos LLM, o texto é truncado em 3.000 caracteres. Os exemplos são selecionados de artigos com alta confiança na classificação por palavras-chave que não pertençam ao gabarito, com verificação anti-vazamento antes de cada chamada à API. O modelo é instruído a responder com o nome exato de uma das 21 especialidades canônicas.

2.3 Métricas de Avaliação

As predições de cada método são comparadas ao gabarito de 178 artigos, calculando-se acurácia, precisão e recall macro e ponderados, e F1 Macro e ponderado. O F1 Macro é a métrica principal, por tratar igualmente todas as 21 classes independentemente do tamanho. São calculados também o Kappa de Cohen e o MCC, adequado a datasets desbalanceados, com intervalos de confiança bootstrap (95%, 1.000 reamostras) para avaliar a estabilidade. A análise de eficiência relativa normaliza o ganho de F1 Macro pelo custo aproximado em tokens, permitindo comparar o retorno de cada estratégia de prompt.

RESULTADOS E DISCUSSÃO

3.1 Desempenho Global

A Tabela 1 apresenta as métricas de desempenho dos seis métodos. O baseline obteve F1 Macro de 0,106 e acurácia de 19,1% — desempenho próximo ao aleatório, confirmado pelo Kappa de 0,121 (concordância insignificante). Todos os métodos LLM superaram expressivamente o baseline, com ganho de ~5× no F1 Macro (Figura 1).

Entre os métodos LLM, o desempenho é próximo: F1 Macro de 0,556 (zero-shot) a 0,567 (few-shot texto). O one-shot com texto completo obteve a maior acurácia (59,3%) e o melhor F1 Ponderado (0,649), enquanto o few-shot com texto completo alcançou o maior F1 Macro. O zero-shot, sem qualquer exemplo, já atinge F1 Macro de 0,556 — apenas 0,011 abaixo do melhor método — com custo 15× inferior.

Tabela 1. Comparativo de métricas por método de classificação.

Método	N	Acurácia	F1 Macro	F1 Pond.	Kappa (κ)
Palavras-chave	178	0,191	0,106	0,191	0,121
Zero-Shot	172	0,564	0,556	0,636	0,510
One-Shot (Texto)	172	0,593	0,562	0,649	0,539

One-Shot (Título)	172	0,552	0,566	0,611	0,501
Few-Shot (Texto)	172	0,564	0,567*	0,632	0,512
Few-Shot (Título)	172	0,587	0,566	0,639	0,532

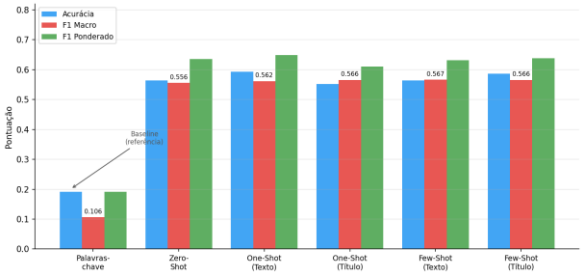

Figura 1 –Acurácia, F1 Macro e F1 Ponderado por método.

3.2 Análise Estatística Avançada

O Kappa de Cohen e o MCC (Figura 2) confirmam que todos os métodos LLM atingem concordância moderada com o gabarito (κ: 0,501–0,539; MCC: 0,522–0,553), enquanto o baseline apresenta κ = 0,121 (insignificante) e MCC = 0,134. As linhas tracejadas indicam limiares de interpretação do Kappa (fraca, moderada, boa).

A concordância entre pares de métodos LLM é alta: Kappa de 0,763 (Zero-Shot vs. One-Shot Título) a 0,900 (One-Shot Texto vs. Few-Shot Título), reforçando que os erros são sistêmicos e independem da estratégia de prompt.

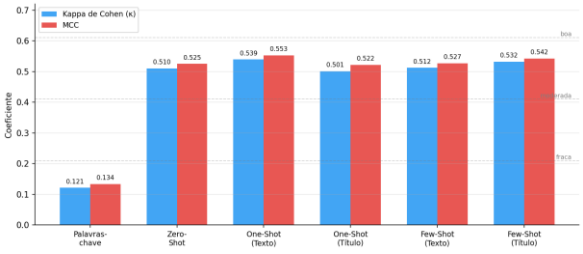

Figura 2 – Cohen's Kappa (κ) e MCC por método.

3.3 Análise por Especialidade

A Figura 3 exibe o desempenho por especialidade no melhor método (Few-Shot Texto), a linha tracejada está centrada em F1 = 0,5, o número de artigos no gabarito é representado por n. Quatro especialidades alcançaram F1 = 1,0: Fonoaudiologia, Odontologia, Oftalmologia e Oncologia. Nefrologia (F1 = 0,968) e Cirurgia (0,875) também apresentaram excelente desempenho. Cardiologia, apesar de ser a segunda especialidade mais frequente no gabarito (n=31), obteve F1 = 0,680, com *recall* de apenas 0,548 indicando que 14 dos 31 artigos foram classificados incorretamente.

O principal padrão de confusão é entre Epidemiologia e Saúde Pública (13 erros), com artigos de epidemiologia classificados como Saúde Pública. Isso tem respaldo em Assis et al. [2021]: pelo indicador Lex-BioMed, Epidemiologia (léxico de fronteira) e Saúde Pública (9,17%, a menor lexicalidade biomédica) compartilham vocabulário próximo, ao contrário de especialidades de alta lexicalidade como Cardiologia (25,3%) e Nefrologia. Essa ambiguidade lexical é um limitante estrutural que ultrapassa a estratégia de prompting. Seis especialidades obtiveram F1 = 0 (Clínica Geral, Gastroenterologia, Reumatologia, Saúde Pública, Urologia e Virologia/Infectologia), todas com apenas 1-5 artigos no gabarito. A correlação de Spearman entre suporte e F1 (ρ = 0,401, p = 0,071) não foi significativa, sugerindo que os erros refletem ambiguidade semântica, não apenas escassez de dados.

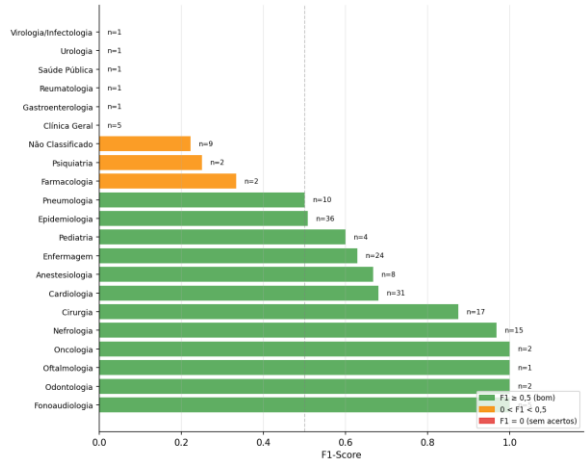

Figura 3 – F1-Score por em Few-Shot (Texto).

3.4 Eficiência e Custo-Benefício

A Tabela 2 e a Figura 4 apresentam a análise de eficiência relativa. O zero-shot tem a melhor relação F1/Custo (0,556): produz F1 Macro de 0,556 com custo unitário. O few-shot com texto completo — melhor F1 absoluto (0,567) — tem eficiência de apenas 0,038, pois custa 15× mais por ganho marginal de 0,011. O one-shot com título emerge como opção intermediária atrativa: custo 1,5× com ganho de 0,010 em F1 Macro e eficiência de 0,377.

Tabela 2 – Eficiência das estratégias de prompt.

Método	F1 Macro	Δ vs Zero-Shot	Custo Relativo	F1/Custo
Zero-Shot	**0,556**	**+0,000**	**1,0×**	**0,556***
One-Shot (Título)	0,566	+0,010	1,5×	0,377
One-Shot (Texto)	0,562	+0,006	6,0×	0,094
Few-Shot (Título)	0,566	+0,011	3,5×	0,162
Few-Shot (Texto)	0,567	+0,012	15,0×	0,038

** Melhor eficiência F1/Custo (negrito).*

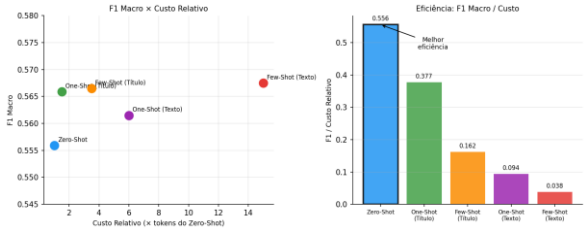

Figura 4 – F1 Macro × Custo Relativo (esq.) e eficiência (razão F1/Custo) por método (dir.).

3.5 Discussão

Os resultados confirmam a primeira questão de pesquisa: o prompting com LLMs supera expressivamente o baseline por palavras-chave na classificação de artigos COVID-19 em português, com ganho de ~5× no F1 Macro. O Kappa de Cohen dos métodos LLM (0,50–0,54) indica concordância moderada, nível razoável para um problema de 21 classes com forte desbalanceamento.

Quanto à segunda questão, o few-shot com texto completo obteve o maior F1 Macro (0,567), mas o custo 15× superior ao zero-shot, com ganho marginal de apenas 0,011, torna-o pouco eficiente. Isso dialoga com Min et al. [2022], que mostraram que os exemplos no few-shot servem sobretudo para demonstrar o formato de saída, não para transmitir conhecimento novo. Como o Claude Haiku 4.5 já internalizou conhecimento biomédico no pré-treinamento [WEI et al., 2022a], adicionar exemplos gera ganho marginal.

A alta concordância inter-método (Kappa 0,76–0,90) indica erros sistêmicos, independentes da estratégia de prompt. Assis et al. [2021] oferecem explicação estrutural: o indicador Lex-BioMed mostra que especialidades como Epidemiologia, Saúde Pública, Clínica Geral e Enfermagem têm baixa lexicalidade biomédica (vocabulário genérico e fronteiriço), enquanto Cardiologia, Nefrologia e Cirurgia têm alta lexicalidade. Os LLMs classificam melhor o segundo grupo (F1 ≥ 0,68), por seu léxico mais distintivo. Prompts com exemplos de contraste explícito entre pares de baixa Lex-BioMed são, assim, o principal caminho para superar esse gargalo.

As principais limitações são: (i) gabarito reduzido (178 artigos) com 21 classes, 8 delas com apenas 1-2 exemplos, tornando as métricas macro instáveis para classes raras; (ii) uso de um único modelo (Claude Haiku 4.5), sem generalização para outros LLMs; (iii) avaliação restrita ao gabarito, sem evidência direta sobre o corpus completo (1.903 artigos).

CONCLUSÃO

Este trabalho demonstrou que estratégias de prompting com LLMs superam expressivamente a classificação por palavras-chave para categorização de artigos biomédicos COVID-19 em português por especialidade médica (F1 Macro: 0,106 → 0,567, ganho ~5×).

Quanto à estratégia ótima, o few-shot com texto completo obteve o melhor F1 Macro absoluto (0,567), mas o zero-shot representa o melhor equilíbrio entre desempenho e custo (F1/Custo = 0,556) sendo a recomendação padrão para pipelines de produção. O one-shot com título configura-se como opção intermediária eficiente (F1/Custo = 0,377, custo 1,5×).

Como trabalhos futuros recomenda-se: (1) ampliar o gabarito de 178 para 500+ artigos, especialmente nas especialidades com 1-2 exemplos; (2) implementar *ensemble* por votação majoritária entre 2-3 métodos; (3) ajustar prompts com exemplos de contraste para pares confusos (Epidemiologia/Saúde Pública); (4) avaliar modelos maiores nas especialidades críticas; e (5) aplicar validação cruzada *k-fold* com conjunto de dados expandido para estimativas mais robustas; (6) Implementação do teste de McNemar [McNEMAR, 1947] que verifica se pares de métodos diferem estatisticamente no padrão de erros.

REFERÊNCIAS

ANTUNES, R. et al. Benchmarking for biomedical natural language processing tasks with a domain specific ALBERT. BMC Bioinformatics, 23, 1-15, 2022.

ASSIS, K. S. P. et al. Lexicalidade biomédica e sua mensuração em um corpus sobre COVID-19 em língua portuguesa. SIMPÓSIO BRASILEIRO DE TECNOLOGIA DA INFORMAÇÃO E DA LINGUAGEM HUMANA, 39-46, 2021.

BROWN, T. B. et al. Language models are few-shot learners. ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS, 1877-1901, 2020.

COHEN, J. A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 20, 37-46, 1960.

DEVLIN, J. et al. BERT: pre-training of deep bidirectional transformers for language understanding. CONFERENCE OF THE NORTH AMERICAN CHAPTER OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, 4171-4186, 2019.

ESTEVA, A. et al. A guide to deep learning in healthcare. Nature Medicine, 25, 24-29, 2019.

GU, Y. et al. Domain-specific language model pretraining for biomedical natural language processing. ACM Transactions on Computing for Healthcare, 3, 1-23, 2021.

LEE, J. et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics, 36, 1234-1240, 2020.

LEITE, J. S.; TAKAHATA, A. K.; STEINBERGER-ELIAS, M. Elaboração de corpus biomédico em Português sobre o Covid-19. Journal of Health Informatics, 242-247, 2020.

McNEMAR, Q. Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika, 12, 153-157, 1947.

MIN, S. et al. Rethinking the role of demonstrations: what makes in-context learning work? CONFERENCE ON EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING, 11048-11064, 2022.

ROBERTS, K. et al. Searching for scientific evidence in a pandemic: an overview of TREC-COVID. Journal of the American Medical Informatics Association, 28, 724-737, 2021.

SOKOLOVA, M.; LAPALME, G. A systematic analysis of performance measures for classification tasks. Information Processing & Management, 45, 427-437, 2009.

SOUZA, F.; NOGUEIRA, R.; LOTUFO, R. BERTimbau: pretrained BERT models for Brazilian Portuguese. BRAZILIAN CONFERENCE ON INTELLIGENT SYSTEMS, 403-417, 2020.

WANG, L. L. et al. CORD-19: the COVID-19 open research dataset. WORKSHOP ON NLP FOR COVID-19 AT ACL, 2020.

WEI, J. et al. Emergent abilities of large language models. Transactions on Machine Learning Research, 2022a.

WEI, J. et al. Finetuned language models are zero-shot learners. INTERNATIONAL CONFERENCE ON LEARNING REPRESENTATIONS, 2022b.

ZHOU, Y. et al. Large language models are human-level prompt engineers. INTERNATIONAL CONFERENCE ON LEARNING REPRESENTATIONS, 2023.