PROPOR 2026

The 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)

Proceedings - Vol. 2 (Demo and Industry Tracks, Best Dissertations, Workshops, and Tutorials)

April 13-16, 2026

The PROPOR organizers gratefully acknowledge the support from the following sponsors.

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
317 Sidney Baker St. S
Suite 400 - 134
Kerrville, TX 78028
USA
Tel: +1-855-225-1962
acl@aclweb.org

ISBN 979-8-89176-387-6

Introduction

These Proceedings cover the 17th edition of the International Conference on the Computational Processing of Portuguese (PROPOR 2026), held in Salvador, Bahia, Brazil, April 13–16, 2026.
PROPOR is a major forum bringing together researchers, developers, and practitioners dedicated to the computational processing of Portuguese and, since the last edition in 2024, Galician. PROPOR promotes the exchange of experiences in the development of methodologies, language resources, tools, applications, and innovative projects covering the target languages. Established as the main scientific event in the area of natural language processing focused on theoretical and technological issues of written and spoken Portuguese and Galician, PROPOR 2026 also highlights the cultural and linguistic diversity present in Salvador, a place where Portuguese has historically interacted with indigenous and African languages, profoundly influencing Brazilian Portuguese and the surrounding culture.
This year's edition of PROPOR includes three tracks: General, Industry, and Demo, as well as three workshops, one tutorial, best MSc and PhD dissertation awards, and three invited speakers. In this edition, papers could be submitted in English or Portuguese.
The scope of PROPOR 2026 is reflected in the organization of the Proceedings, which are divided into two volumes. The first volume presents the accepted papers of the General track. The present (second) volume is dedicated to the Demo and Industry tracks, the Best Dissertation awards, and the workshops co-located with PROPOR 2026, namely:

- The First Workshop on Language Technologies for Health (Lang4Health);
- The Third Student Research Workshop (SRW);
- The Fourth Workshop on Digital Humanities and Natural Language Processing (DHandNLP).

This edition of PROPOR 2026 also features the following tutorial:

- *From Syntax to Semantics: Introducing UMR for NLP Annotation*
  Adriana S. Pagano (UFMG), Magali Sanches Duran (USP), Federica Gamba (CUni)

PROPOR 2026 is structured as a four-day event, commencing with a full day of workshops and tutorials, followed by three days of communications, posters, demos, and community meetings. Our sincere thanks go to every person and institution involved in the complex organization of this event, especially the members of the Program Committee, the Workshop and Tutorial chairs, the chairs of the Demo and Industry tracks, the Best Dissertation chairs and jury, the invited speakers, and the general organization staff. We are also grateful to the agencies and organizations that supported and promoted the conference. Thank you all for participating and contributing to the success of this conference!

Marlo Souza and Iria de-Dios-Flores, General Chairs
Diana Santos and Larissa Freitas, Program Chairs
Jackson Wilke da Cruz Sousa and Eugénio Ribeiro, Editorial Chairs

Organizing Committee

General Chairs

Marlo Souza, Universidade Federal da Bahia, Brazil
Iria de-Dios-Flores, Universitat Pompeu Fabra - Barcelona, Spain

Program Chairs

Diana Santos, Universitetet i Oslo, Norway
Larissa Freitas, Universidade Federal de Pelotas, Brazil

Editorial Chairs

Jackson Wilke da Cruz Souza, Universidade Federal da Bahia, Brazil
Eugénio Ribeiro, Iscte-IUL & INESC-ID Lisboa, Portugal

Demo Chairs

Evandro Fonseca, Blip, Brazil
Susana Sotelo, Universidade de Santiago de Compostela, Spain

Industry Track Chairs

Clarissa Xavier, Banrisul, Brazil
Henrico Brum, Sinch AB, Sweden

Best Dissertation Chairs

Marcos Garcia, Universidade de Santiago de Compostela, Spain
Aline Paes, Universidade Federal Fluminense, Brazil

Workshops and Tutorial Chairs

Roney Lira de Sales Santos, Universidade Federal da Bahia, Brazil
Renata Vieira, Universidade de Évora, Portugal

Workshop on Language Technologies for Health (Lang4Health)

Aline Villavicencio, University of Exeter, UK
Rodrigo Wilkens, University of Exeter, UK
Helena Caseli, Federal University of São Carlos, Brazil
Vânia Neris, Federal University of São Carlos, Brazil

Student Research Workshop (SRW)

Livy Real, Universidade Federal do Amazonas, Brazil
Jackson Wilke da Cruz Souza, Universidade Federal da Bahia, Brazil

Workshop on Digital Humanities and Natural Language Processing (DHandNLP)

Leonardo Zilio, Université Catholique de Louvain, CENTAL, LSTI, Belgium
Helena Freire Cameron, Instituto Politécnico de Portalegre, CIDEHUS, Portugal
Maria José B. Finatto, Universidade Federal do Rio Grande do Sul, CNPq/PPG-LETRAS, Brazil
Renata Vieira, Évora University, CIDEHUS, Portugal

Local Organization

Marlo Souza, Universidade Federal da Bahia, Brazil
Daniela Barreiro Claro, Universidade Federal da Bahia, Brazil
Jackson Wilke da Cruz Souza, Universidade Federal da Bahia, Brazil
Lilian Teixeira, Universidade Federal da Bahia, Brazil
Rerisson Cavalcante, Universidade Federal da Bahia, Brazil
Robespierre Pita, Universidade Federal da Bahia, Brazil
Roney Lira de Sales Santos, Universidade Federal da Bahia, Brazil

Program Committee

Demo Track

Andre Carvalho, Universidade Federal do Amazonas, Brazil
Bruno Souza Cabral, Escavador, Brazil
Daniela Schmidt, Universidade de Évora, Portugal
Jesus M. Benitez Baleato, Universidade de Santiago de Compostela, Spain
José Ramom Pichel, imaxin.software & Universidade de Santiago de Compostela, Spain
Livy Real, Universidade Federal do Amazonas, Brazil
Luis Trigo, Universidade do Porto, Portugal
Luiz Merschmann, Universidade Federal de Lavras, Brazil
Saullo Oliveira, Pontifícia Universidade Católica de Campinas, Brazil
Thiago Pardo, Universidade de São Paulo, Brazil

Industry Track

Beatriz Fagundes, Clio, Canada
Fabio Rezende de Souza, University of São Paulo, Brazil
Marcio Bigolin, Instituto Federal do Rio Grande do Sul, Brazil
Nataly Leopoldina Patti da Silva, SiDi, Brazil
Sidney Evaldo Leal, Venturus, Brazil
Vítor Rodrigues Tonon, Universidade Estadual do Norte do Paraná, Brazil

Best Dissertation

Alberto Abad, Instituto Superior Técnico, Universidade de Lisboa & INESC-ID Lisboa, Portugal
Aline Vanin, Federal University of Health Sciences of Porto Alegre, Brazil
Amália Mendes, Universidade de Lisboa, Portugal
Arnaldo Candido Junior, Universidade Estadual Paulista, Brazil
Daniela Barreiro Claro, Universidade Federal da Bahia, Brazil
David Vilares, Universidade da Coruña, Spain
Helena Caseli, Federal University of São Carlos, Brazil
Ivandre Paraboni, University of São Paulo, Brazil
Jorge Baptista, Universidade do Algarve & INESC-ID Lisboa, Portugal
Luís M. S. Gomes, Faculdade de Ciências da Universidade de Lisboa, Portugal
Maria José Finatto, Universidade Federal do Rio Grande do Sul, Brazil
Magali S. Duran, University of São Paulo, Brazil
Marcelo Finger, University of São Paulo, Brazil
Marcos Fernandez-Pichel, Universidade de Santiago de Compostela, Spain
Maria das Graças Volpe Nunes, University of São Paulo, Brazil
Pablo Gamallo, Universidade de Santiago de Compostela, Spain
Paulo Quaresma, Universidade de Évora, Portugal
Plinio A. Barbosa, Universidade Estadual de Campinas, Brazil
Renata Vieira, Évora University, CIDEHUS, Portugal
Ricardo Ribeiro, Iscte-IUL & INESC-ID Lisboa, Portugal
Vladia C. M. Pinheiro, Universidade de Fortaleza, Brazil

Workshop on Language Technologies for Health (Lang4Health)

Aline Paes, Institute of Computing / Universidade Federal Fluminense, Brazil
Ana Cleide Guimbal de Aquino, Universidade Federal Rural da Amazônia, Brazil
César Sperb, Federal University of Pelotas, Brazil
Claudia Moro, Pontifícia Universidade Católica do Paraná, Brazil
Elisa Terumi Rubel Schneider, Pontifícia Universidade Católica do Paraná, Brazil
Eloize Seno, Federal Institute of São Paulo, Brazil
Emerson Paraiso, Pontifícia Universidade Católica do Paraná, Brazil
Helena Caseli, Federal University of São Carlos, Brazil
Ivandre Paraboni, University of São Paulo, Brazil
João Papa, São Paulo State University, Brazil
Luciana Salgados, Universidade Federal Fluminense, Brazil
Marcelo Finger, University of São Paulo, Brazil
Maria José Finatto, Universidade Federal do Rio Grande do Sul, Brazil
Marilia Silveira, Federal University of Pelotas, Brazil
Mateus Monteiro, Brazil
Murilo Vargas da Cunha, Federal University of Pelotas, Brazil
Paula Souza, Universidade Federal de São Carlos, Brazil
Renata Vieira, Évora University, CIDEHUS, Portugal
Renato Silva, University of São Paulo, Brazil
Rodrigo Wilkens, University of Exeter, UK
Sandra Rodrigues, Universidade Federal de Lavras, Brazil
Tiago Torrent, Universidade Federal de Juiz de Fora, Brazil

Student Research Workshop (SRW)

Ana Mata, Universidade de Lisboa, Portugal
António Branco, Universidade de Lisboa, Portugal
Arnaldo Candido Junior, Universidade Estadual Paulista, Brazil
Eulanda M. dos Santos, Universidade Federal do Amazonas, Brazil
Evelin Amorim, INESC-TEC, Portugal
Fábio Lobato, Universidade do Oeste do Pará, Brazil
Fernanda Lopez-Escobedo, Universidad Nacional Autónoma de México, Mexico
Jackson Wilke da Cruz Souza, Universidade Federal da Bahia, Brazil
Jorge Baptista, University of Algarve & INESC-ID Lisboa, Portugal
Leo Sampaio Ferraz Ribeiro, Universidade de São Paulo, Brazil
Leonel Figueiredo de Alencar, Universidade Federal do Ceará, Brazil
Nádia Silva, Universidade Federal de Goiás, Brazil
Raquel Freitag, Universidade Federal de Sergipe, Brazil
Saullo Oliveira, Pontifícia Universidade Católica de Campinas, Brazil
Thiago Pardo, Universidade de São Paulo, Brazil

Workshop on Digital Humanities and Natural Language Processing (DHandNLP)

Álvaro Iriarte, University of Minho, Portugal
Ana Ribeiro, University of Évora, Portugal
Bruno Maroneze, Federal University of Grande Dourados, Brazil
Cassia Trojahn, University of Toulouse, France
Fernanda Olival, University of Évora, Portugal
Idalete da Silva Dias, University of Minho, Portugal
Marcus Dores, Federal University of Bahia, Brazil

Paulo Quaresma, University of Évora, Portugal
Raquel Amaro, NOVA University, Portugal
Sandro Marengo, Federal University of Sergipe, Brazil
Silvana Silva, Federal University of Rio Grande do Sul, Brazil

Table of Contents

ARAMIS: Uma ferramenta web integrada com LLM open-source para apoio à correção de TCCs de estudantes de graduação

Gustavo Campelo de Sousa[1], Pablo Kauan Martins Timbó[1], Luiz Zairo Bastos Viana[1], Antônio Emerson Barros Tomaz[2], José Wellington Franco da Silva[1], Carlos de Oliveira Caminha[1]
[1]Kunumi Lab – Universidade Federal do Ceará (UFC), Fortaleza – CE
[2]GSIPP – Universidade Federal do Ceará (UFC), Crateús – CE
gustavocraft321@gmail.com, pablokauan@alu.ufc.br, zairobastos@gmail.com, emerson@crateus.ufc.br, wellington@crateus.ufc.br, caminha@ufc.br,

Abstract

A revisão de trabalhos acadêmicos é uma etapa crucial, porém onerosa, na formação de pesquisadores. Trabalhos anteriores obtiveram bons resultados com abordagens automatizadas de revisão em inglês. Nesse contexto, apresentamos o ARAMIS [1], uma ferramenta multiagente baseada em *Large Language Models* (LLM) *open-source*, projetada para revisar Trabalhos de Conclusão de Curso (TCC) em português. A solução foca em três pilares: correção gramatical, encadeamento lógico e rigor metodológico, permitindo ao usuário receber revisões estruturadas para cada pilar escolhido. Mesmo em estágio experimental, os testes atingiram ótimos resultados de usabilidade ao aplicar o *System Usability Scale* (SUS), obtendo uma pontuação de 90,5/100.

1 Introdução

O Trabalho de Conclusão de Curso (TCC) constitui uma etapa importante na formação acadêmica, por representar a consolidação dos conhecimentos desenvolvidos ao longo da graduação e por exigir do estudante domínio técnico, clareza na escrita e organização argumentativa (Guedes and Guedes, 2012). De acordo com a Associação Brasileira de Normas Técnicas (2024), trata-se de um documento elaborado segundo princípios e normas específicas, com a finalidade de obter o grau de bacharelado ou licenciatura. Nesse contexto, a produção do TCC demanda atenção contínua, especialmente porque falhas de escrita e de estrutura comprometem a qualidade do texto e dificultam a comunicação científica (Carboni and Nogueira, 2008). Entre os problemas frequentemente observados em produções acadêmicas, destacam-se erros de grafia, inadequações de pontuação e fragilidades na construção textual (Lunsford and Lunsford, 2008).

Com o avanço das tecnologias de Processamento de Linguagem Natural (PLN), diferentes abordagens de correção automatizada passaram a ser empregadas para apoiar a escrita acadêmica, reduzindo erros tipográficos e gramaticais e melhorando a experiência do usuário (Srivarsha et al., 2025). Mais recentemente, o surgimento dos *Large Language Models* (LLMs) ampliou de forma significativa esse cenário, possibilitando análises mais refinadas e contextualizadas de textos científicos (Liang et al., 2024). Trabalhos recentes têm demonstrado o potencial desses modelos para gerar revisões precisas e direcionadas, sobretudo em língua inglesa, seja com modelos proprietários (Chamoun et al., 2024; D'Arcy et al., 2024), seja com alternativas de código aberto (Idahl and Ahmadi, 2025).

Além das aplicações em linguagem natural, os LLMs vêm demonstrando elevada capacidade de adaptação a diferentes domínios, consolidando-se como uma tecnologia de propósito geral. Esses modelos têm sido utilizados em áreas diversas, como visão computacional (Wang et al., 2024), geração de código (Gu, 2023) e previsão de séries temporais (Bastos et al., 2025b). Esse avanço evidencia que os LLMs podem servir de base para o desenvolvimento de ferramentas especializadas, voltadas não apenas à geração de texto, mas também ao suporte à tomada de decisão, à automação de tarefas e à análise de conteúdos técnicos.

Esse movimento já pode ser observado em aplicações construídas para finalidades específicas. No domínio de séries temporais, por exemplo, soluções como o *LLM4Time* (Bastos et al., 2025a) exploram o uso de LLMs para previsão, enquanto bibliotecas como o *LLM4Series* (Silva et al., 2026) auxiliam na construção de *prompts* e oferecem suporte a diferentes *back-ends* para tarefas preditivas. Em outro contexto, ferramentas como o *OBI-UAN* (Bastos et al., 2025c) utilizam esses modelos para apoiar a explicação de questões de programação. Apesar desse avanço, ainda há uma lacuna no suporte à língua portuguesa, especialmente no que se refere à revisão de textos acadêmicos com foco não apenas

[1]https://youtu.be/lw2GSQ7aiW0

Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)
Vol. 2, pages 1–4
April 13-16, 2026

em aspectos gramaticais, mas também em elementos mais amplos, como coerência argumentativa e rigor metodológico.

Diante dessa lacuna, este trabalho apresenta o ARAMIS (*Academic Review Agents for Methodological Improvements*), uma ferramenta *open-source* voltada à revisão de TCCs em português, fundamentada em Grandes Modelos de Linguagem. A proposta integra um LLM *open-source* a uma arquitetura orientada por *prompts* especializados, capaz de fornecer *feedback* linguístico e estrutural ao estudante. Com isso, busca-se contribuir para a melhoria da qualidade textual e do rigor científico das produções acadêmicas, além de tornar mais eficiente o processo de acompanhamento e orientação.

As seções a seguir apresentam a descrição do sistema, a estratégia de avaliação adotada, os resultados obtidos e as perspectivas de continuidade.

2 Descrição da Ferramenta

O ARAMIS[2] é uma ferramenta *Web* baseada em uma arquitetura multiagente, integrada com o modelo *open-source gpt-oss-20b*[3], pela API do *Hugging Face*. É composta por três agentes: correção gramatical, encadeamento lógico e rigor metodológico. Voltada para alunos de graduação, o ARAMIS tem o intuito de fornecer *feedback* direcionado sobre os elementos do texto inserido. O usuário deve acessar a seção "Nova Correção" da Figura 1 e, nela, preencher algumas informações de pré-configuração disponíveis na interface, além de inserir o texto que receberá a revisão. Uma vez inserido, a geração da revisão inicia-se ao clicar no botão "Iniciar Análise" e, após um curto tempo de espera, o *feedback* é gerado e exibido ao usuário em uma nova seção chamada "Análises", conforme exibido na Figura 2; tal seção exibe para o usuário todas as revisões geradas, permitindo que ele as consulte em qualquer momento oportuno. Outras funcionalidades principais disponíveis são o *login*, o cadastro e a seção de histórico de análises, presentes na seção "Minhas Análises".

2.1 Arquitetura

O ARAMIS é composto por uma arquitetura baseada em multiagentes, na qual cada agente é integrado ao modelo *open-source gpt-oss-20b* e orquestrado pelo *framework Agno*[4]. A Figura 3 exibe o fluxograma da ferramenta, ilustrando, desde a entrada do texto, as etapas de pré-configuração e de preenchimento de *prompts*, a definição da requisição, e segue para a orquestração dos agentes especializados (correção gramatical, encadeamento lógico e rigor metodológico), até a organização das análises e a geração do *review* final consolidado.

Figure 1: Interface da seção "Nova Correção".

Figure 2: Interface da seção "Análises Realizadas".

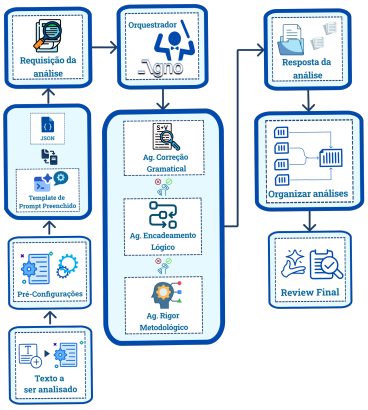


Figure 3: Fluxograma da arquitetura do ARAMIS.

[2] http://enginelab.ufc.br/aramis
[3] https://huggingface.co/openai/gpt-oss-20b
[4] https://docs.agno.com/

2.2 Agentes do ARAMIS

Em sua versão atual, o ARAMIS integra três agentes com características próprias:

- **Correção Gramatical**: Agente voltado para identificar e listar erros ortográficos e equívocos de acentuação;
- **Encadeamento Lógico**: Agente preparado para lidar com a coerência entre sentenças;
- **Rigor Metodológico**: Agente concentrado em verificar a coerência entre o problema de pesquisa, os objetivos, o método e a análise dos resultados.

2.3 Modelagem dos Prompts

Os *prompts* foram estruturados de modo a receberem orientações explícitas sobre os detalhes e o formato da resposta a serem retornados. Além disso, há *placeholders* que o sistema preenche automaticamente com as informações inseridas pelo usuário na etapa anterior à ativação da revisão. Para a modelagem dos *prompts*, utilizou-se o *few-shot prompting* (IBM, 2025). Abaixo, detalhamos quais variáveis são substituídas:

- **secao_desejada**: Indica nos *prompts* a seção que será abrangida para análise;
- **titulo_tcc**: Contém o título do trabalho do aluno;
- **area_conhecimento_tcc**: Representa a área do conhecimento a qual o TCC está atrelado;
- **nivel_rigor_modelo**: Define o nível de rigor com que o modelo deve avaliar o trecho do TCC do aluno.

3 Avaliação e Discussão

A usabilidade do sistema foi avaliada por meio do *System Usability Scale* (SUS) (Brooke, 1995). O SUS é uma métrica que mensura a usabilidade subjetiva, com 10 itens em uma escala *Likert* de 5 pontos, que varia de "Discordo Completamente" (1) a "Concordo Completamente" (5) (Joshi et al., 2015).

O *score* final, conforme a Fórmula 1 do SUS, é calculado pela soma das contribuições individuais de cada item do questionário. Para os itens de numeração ímpar, formulados de maneira positiva, subtrai-se 1 do valor atribuído pelo respondente. Para os itens de numeração par, formulados de maneira negativa, a contribuição é obtida subtraindo-se a resposta de 5.

$$SUS = 2,5 \times \left[\sum_{i=1}^{10} f(x_i) \right] \quad (1)$$

Embora a proposta do SUS não defina categorias qualitativas para a interpretação dos *scores*, este trabalho adota a escala sugerida por Bangor et al. (2009), amplamente utilizada. Essa classificação permite qualificar o desempenho do sistema conforme os intervalos:

- **Excelente**: Pontuação entre 90 e 100;
- **Bom**: Pontuação entre 80 e 89;
- **Aceitável**: Pontuação entre 70 e 79;
- **Precisa Melhorar**: Pontuação entre 60 e 69;
- **Ruim**: Pontuação inferior a 60.

O *feedback* dos usuários indicou a simplicidade de uso, a boa construção das seções, a qualidade e a precisão das revisões dos agentes, especialmente as do agente de correção gramatical, além do bom desempenho do LLM integrado aos agentes, que rapidamente gerou as revisões e identificou corretamente as incoerências, gerando *feedbacks* úteis. Alguns participantes relataram que os comentários retornados pelo LLM são fundamentais para que o usuário possa obter *feedback* sobre o andamento do trabalho e identificar em que precisa melhorar. Por fim, os usuários relataram a intenção de continuar utilizando o ARAMIS para revisar seus trabalhos, sem se limitar aos testes realizados anteriormente.

Apesar dos resultados promissores, este trabalho apresenta limitações relacionadas ao número reduzido de modelos analisados, bem como à quantidade restrita de participantes, composta por dez usuários de uma única instituição, o que pode comprometer a generalização dos resultados. Somam-se a isso as restrições impostas pelo uso de APIs gratuitas, que têm um limite de requisições, além das limitações da infraestrutura de hospedagem utilizada, que por não suportar localmente modelos como o *gpt-oss-20b*, tornou-se necessária a utilização da API externa do *Hugging Face*.

4 Trabalhos Futuros

O sistema proposto está em fase de melhoria contínua, com esforços para aprimorar a usabilidade e a qualidade das revisões geradas pelo modelo *open-source*. Trabalhos futuros pretendem aplicar técnicas comparativas para melhor seleção do modelo integrado à ferramenta. Pretende-se ampliar a base de usuários para realizar a avaliação da usabilidade da ferramenta. Sugere-se o uso de *fine-tuning*

em um modelo *open-source* para maior direcionamento e precisão nas revisões, além de avaliar também a qualidade das revisões geradas. Além disso, planeja-se hospedar o sistema em um servidor que suporte um modelo como o *gpt-oss-20b*, para que não dependa mais de APIs externas.

Agradecimentos

Agradecemos ao Instituto Kunumi pelo financiamento, pela infraestrutura computacional disponibilizada e pelo apoio à realização e apresentação deste trabalho.

Referências

Associação Brasileira de Normas Técnicas. 2024. NBR 14724: Informação e documentação – Trabalhos acadêmicos – Apresentação. Rio de Janeiro: ABNT. 4. ed.

Aaron Bangor, Philip Kortum, and James Miller. 2009. Determining what individual sus scores mean: Adding an adjective rating scale. *Journal of usability studies*, 4(3):114–123.

Zairo Bastos, Carlos Caminha, and Wellington Franco. 2025a. Llm4time: Uma ferramenta interativa para previsão de séries temporais com modelos largos de linguagem. In *Anais Estendidos do XL Simpósio Brasileiro de Bancos de Dados*, pages 100–105, Porto Alegre, RS, Brasil. SBC.

Zairo Bastos, João David Freitas, José Wellington Franco da Silva, and Carlos Caminha. 2025b. Prompt-driven time series forecasting with large language models. In *ICEIS (1)*, pages 309–316.

Zairo Bastos, Raylander Marques, Gabriel Rudan, Marlon Duarte, and Wellington Franco. 2025c. Obi-uan: Um agente para auxiliar nos estudos da olimpíada brasileira de informática. In *Anais Estendidos do XL Simpósio Brasileiro de Bancos de Dados*, pages 94–99, Porto Alegre, RS, Brasil. SBC.

John Brooke. 1995. Sus: A quick and dirty usability scale. *Usability Eval. Ind.*, 189.

Rosadélia Malheiros Carboni and Valnice de Oliveira Nogueira. 2008. Facilidades e dificuldades na elaboração de trabalhos de conclusão de curso. *ConScientiae Saúde*, 3:65–72.

Eric Chamoun, Michael Schlichtkrull, and Andreas Vlachos. 2024. Automated focused feedback generation for scientific writing assistance. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9742–9763, Bangkok, Thailand. Association for Computational Linguistics.

Mike D'Arcy, Tom Hope, Larry Birnbaum, and Doug Downey. 2024. Marg: Multi-agent review generation for scientific papers. *Preprint*, arXiv:2401.04259.

Qiuhan Gu. 2023. Llm-based code generation method for golang compiler testing. In *Proceedings of the 31st ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, pages 2201–2203.

Hermila Tavares Vilar Guedes and Jorge Carvalho Guedes. 2012. Avaliação, pelos estudantes, da atividade" trabalho de conclusão de curso" como integralização do eixo curricular de iniciação à pesquisa científica em um curso de medicina. *Revista Brasileira de Educação Médica*, 36(2):162–171.

IBM. 2025. What is few-shot prompting? https://www.ibm.com/think/topics/few-shot-prompting. Acesso em: 24 ago. 2025.

Maximilian Idahl and Zahra Ahmadi. 2025. OpenReviewer: A specialized large language model for generating critical scientific paper reviews. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (System Demonstrations)*, pages 550–562, Albuquerque, New Mexico. Association for Computational Linguistics.

Ankur Joshi, Saket Kale, Satish Chandel, and D Kumar Pal. 2015. Likert scale: Explored and explained. *British journal of applied science & technology*, 7(4):396.

Weixin Liang, Yuhui Zhang, Hancheng Cao, Binglu Wang, Daisy Yi Ding, Xinyu Yang, Kailas Vodrahalli, Siyu He, Daniel Scott Smith, Yian Yin, and 1 others. 2024. Can large language models provide useful feedback on research papers? a large-scale empirical analysis. *NEJM AI*, 1(8):AIoa2400196.

Andrea A. Lunsford and Karen J. Lunsford. 2008. "mistakes are a fact of life": A national comparative study. *College Composition and Communication*, 59(4):781–806.

Wesley Barbosa Silva, Maria Fernanda Aquino Freitas Scarcela, Luiz Zairo Bastos Viana, Carlos Caminha, João Paulo do Vale Madeiro, and José Wellington Franco da Silva. 2026. LLM4series: Structured prompting for time series forecasting with LLMs. In *1st ICLR Workshop on Time Series in the Age of Large Models*.

Nalla Srivarsha, Gummadi Nithin, Shanmugasundaram Hariharan, Bibhuti Bhusan Dash, Subrata Chowdhury, and Sudhansu Shekhar Patra. 2025. Auto text correction using nlp techniques. In *2025 6th International Conference on Mobile Computing and Sustainable Informatics (ICMCSI)*, pages 590–594. IEEE.

Wenhai Wang, Zhe Chen, Xiaokang Chen, Jiannan Wu, Xizhou Zhu, Gang Zeng, Ping Luo, Tong Lu, Jie Zhou, Yu Qiao, and 1 others. 2024. Visionllm: Large language model is also an open-ended decoder for vision-centric tasks. *Advances in Neural Information Processing Systems*, 36.

From Complexity Scores to Readable Texts: iREAD4SKILLS for Adult Literacy in Portuguese

Jorge Baptista[1,2], **Eugénio Ribeiro**[1,3], **Nuno Mamede**[1], **David Antunes**[1], **Raquel Amaro**[4]
[1] INESC-ID Lisboa, Portugal
{jorge.baptista,eugenio.ribeiro,nuno.mamede,david.f.l.antunes}@inesc-id.pt
[2] Faculdade de Ciências Humanas e Sociais, Universidade do Algarve, Portugal
[3] Instituto Universitário de Lisboa (ISCTE-IUL), Portugal
[4] CLUNL, Universidade NOVA de Lisboa, Portugal
raquelamaro@fcsh.unl.pt

Abstract

Adult Learning (AL) programmes need short, trustworthy texts that match learners' reading abilities, but educators rarely have time, tools, or evidence-based guidelines to select and adapt materials consistently. We present a live demo of iREAD4SKILLS for European Portuguese: a web-based system that (i) estimates readability/complexity for AL-oriented levels aligned with CEFR, (ii) highlights where complexity concentrates (lexical, grammatical, semantic), and (iii) supports rewriting by offering actionable, level-aware suggestions and curated lexical resources. The demo emphasises transparency and "trainer-first" workflows: users see *why* a text is complex and *how* to revise it without losing meaning.

1 Motivation

Readability assessment has moved well beyond classic formulas, toward feature-rich and neural approaches that can support practical interventions in educational settings (Collins-Thompson, 2014; Crossley et al., 2019; Lee et al., 2021). In Adult Learning, the target is not "school-grade" readability but *functional* comprehension and task-oriented reading under time constraints, often for learners with limited literacy backgrounds (OECD, 2024, 2023, 2013; Steeds, 2001). For European Portuguese, recent work has shown the feasibility of automatic readability modelling with supervised learning and a careful feature design (Ribeiro et al., 2024a,b), while related efforts address generalisation and feature transfer across Portuguese settings (Akef et al., 2024).

Our contribution in this demo paper is the *interactive bridge* from a readability score to concrete, level-driven rewriting: users can (i) paste/upload a text (or even a mobile phone picture of it, to be OCR processed), (ii) obtain an overall level prediction with an interpretable breakdown, and (iii) iteratively revise the same text guided by explanations grounded on project research and resources (Amaro et al., 2025; Monteiro et al., 2025; Blanco Escoda et al., 2023; Amaro et al., 2024). The intended audience includes low-literacy adults, AL trainers, curriculum designers, librarians, publishers, and researchers interested in Portuguese readability and explainable educational NLP.

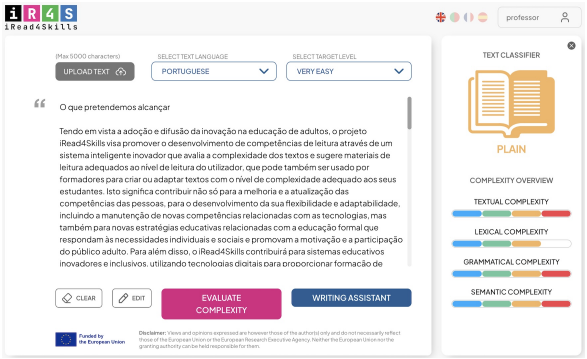


Figure 1: Classifier view: overall level prediction with a compact breakdown by complexity dimension (Portuguese example).

2 iREAD4SKILLS System in one minute

iREAD4SKILLS is an open-access platform that supports multilingual complexity assessment of texts; here we focus on the Portuguese pipeline and AL use cases. A typical interaction is: (i) select language and desired target level (e.g., Portuguese, *Very Easy*), (ii) submit a text (up to a short passage), (iii) inspect the predicted level and the four-dimensional overview (textual, lexical, grammatical, semantic), and (iv) open the writing assistant to see sentence and token-level signals and access targeted resources.

Figure 1 illustrates the *Text Classifier* view, which returns the predicted level and a compact

Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)
Vol. 2, pages 5–7
April 13-16, 2026

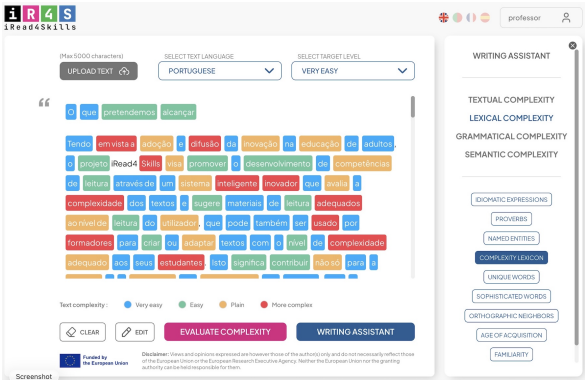


Figure 2: Writing assistant view: token-level highlighting and access to level-oriented resources (e.g., lexicons, unique/sophisticated words).

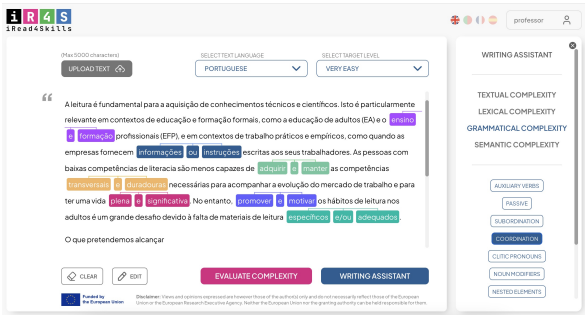


Figure 3: Writing assistant view: sequence-level highlighting coordinated elements.

complexity overview, accross 4 yardsticks: Textual, Lexical, Grammatical, and Semantic Complexity. Figures 2 to Figure 4 show the *Writing Assistant* view, where tokens and sequences, respectively, are highlighted according to their contribution to complexity (e.g., unusually rare items, specialized vocabulary, long-distance syntactic dependencies, or semantic density).

3 Under the hood

The Portuguese pipeline follows a practical design aimed at AL settings: robust text processing and an interpretable feature layer, combined with modern modelling where adopted. At a high level, the system uses (i) curated corpora and level definitions tailored to iRead4Skills, including open-access datasets and lexicons by complexity level (Pintard et al., 2024b,a; Wilkens et al., 2024); (ii) feature sets grounded in readability research (e.g., lexical rarity/frequency, cohesion and discourse signals, and syntactic complexity proxies) (Chen and Meurers, 2018; McNamara et al., 2010; Crossley et al., 2017); and (iii) supervised readability models for European Portuguese aligned with CEFR-inspired

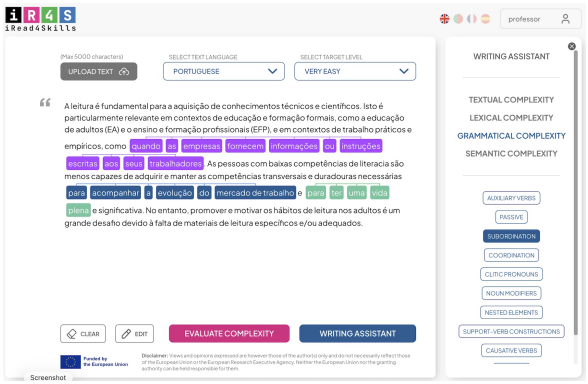


Figure 4: Writing assistant view: sequence-level highlighting with long-distance dependencies (subclauses).

targets (Council of Europe, 2020; Ribeiro et al., 2024a,b). This combination supports not only prediction but also explanation: the interface surfaces *what* is complex and offers *revision directions* consistent with the intended level.

Crucially, the demo is not presented as a black box. Instead, it operationalizes a "trainer-first" loop: (i) detect complexity hotspots, (ii) consult level-specific lexical resources and guidance, (iii) revise, and (iv) re-evaluate—a workflow aligned with educational NLP tools for complexity analysis (Chen and Meurers, 2016) and the broader readability literature (Collins-Thompson, 2014; Crossley et al., 2019).

4 Demo plan

The demo is designed for a fast, hands-on experience and for audience participation. In the session, we will run three short scripts with live interaction: (i) *cold-start diagnosis*—attendees paste a short Portuguese text (e.g., a leaflet, civic information, workplace safety note) and the system returns level + breakdown; and (ii) *guided rewriting*—we open the writing assistant and determine different constructs contributing to the text's complexity. The original text can be rewritten and its complexity reassessed to meet the target user literacy level.

To maximise engagement, we will keep a "challenge bowl" of real-world micro-texts (one per domain: health, public services, work, training materials). Participants can vote for the most effective revision and we re-run the classifier to show before/after deltas. When feasible, we also show cross-language comparability by briefly switching to another supported language, reinforcing the platform's multilingual scope.

5 Impact and availability

By aligning readability assessment with AL practice, iREAD4SKILLS supports: (i) faster selection of appropriate materials for low-literacy adult learners, (ii) consistent adaptation decisions backed by explainable signals, and (iii) resource creation and curation (lexicons and corpora) for the Portuguese community. The platform is openly accessible, and major project resources (datasets, lexicons, and level definitions) are distributed via project website[1] and Zenodo.[2]

Acknowledgments

Work supported by European Comission under Project iRead4Skills, HORIZON-CL2-2022-TRANSFORMATIONS-01-07, DOI: 10.3030/101094837, and by national funds through Fundação para a Ciência e a Tecnologia, I.P. (FCT) under projects UID/50021/2025 (DOI: https://doi.org/10.54499/UID/50021/2025), UID/PRR/50021/2025 (DOI: https://doi.org/10.54499/UID/PRR/50021/2025) and UID/03213/2025 (DOI: https://doi.org/10.54499/UID/03213/2025).

References

S. Akef, A. Mendes, D. Meurers, and P. Rebuschat. 2024. Investigating the Generalizability of Portuguese Readability Assessment Models Trained Using Linguistic Complexity Features. In *PROPOR 2024*, pages 332–341.

R. Amaro, S. Correia, R. Monteiro, A. Pintard, M. Mourinho, and S. Barbosa. 2025. Cadre de référence pour la complexité textuelle destiné aux adultes peu alphabétisés : niveaux et descripteurs dans le cadre du projet iread4skill. *Langues & Parole(s)*, 10:57–119.

Raquel Amaro, Ricardo Monteiro, Thomas François, and J. Nagant de Deuxchaisnes. 2024. iread4skills - data set 2: Annotated corpora report.

Xavier Blanco Escoda, Raquel Amaro, Thomas François, and Marcos Garcia. 2023. iread4skills - baselines for complexity lexicons definition.

X. Chen and D. Meurers. 2016. CTAP: A Web-based Tool Supporting Automatic Complexity Analysis. In *CL4LC*, pages 113–119.

X. Chen and D. Meurers. 2018. Word Frequency and Readability: Predicting the Text-level Readability with a Lexical-level Attribute. *Journal of Research in Reading*, 41(3):486–510.

K. Collins-Thompson. 2014. Computational assessment of text readability: A survey of current and future research. *International Journal of Applied Linguistics*, 165(2):97–135.

Council of Europe. 2020. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment – Companion Volume*. Council of Europe Publishing, Strasbourg.

S. A. Crossley, S. Skalicky, and M. Dascalu. 2019. Moving beyond classic readability formulas: New methods and new models. *Journal of Research in Reading*, 42(3-4):541–561.

S. A. Crossley, S. Skalicky, M. Dascalu, D. S. McNamara, and K. Kyle. 2017. Predicting text comprehension processing and familiarity in adult readers: New approaches to readability formulas. *Discourse Processes*, 54(5-6):340–359.

B. W. Lee, Y. S. Jang, and J. Lee. 2021. Pushing on text readability assessment: A transformer meets handcrafted linguistic features. In *EMNLP 2021*, pages 10669–10686.

D. S. McNamara, M. M. Louwerse, P. M. McCarthy, and A. C. Graesser. 2010. Coh-metrix: Capturing linguistic features of cohesion. *Discourse Processes*, 47(4):292–330.

R. Monteiro, S. Correia, R. Amaro, M. Moutinho, S. Barbosa, and M. L. Reis. 2025. Níveis e descritores de complexidade textual para adultos de baixa literacia: um referencial do projeto iread4skills. *Revista da Associação Portuguesa de Linguística*, 1(13):193–222.

OECD. 2013. *The Survey of Adult Skills Reader's Companion*. OECD Publishing, Paris.

OECD. 2023. PISA 2022 Results (Volume I): The State of Learning and Equity in Education, PISA . Technical report, OECD Publishing, Paris.

OECD. 2024. Do Adults Have the Skills They Need to Thrive in a Changing World?: Survey of Adult Skills 2023. Technical report, OECD Publishing, Paris.

A. Pintard, T. François, J. Nagant de Deuxchaisnes, S. Barbosa, M. L. Reis, M. Moutinho, R. Monteiro, R. Amaro, S. Correia, S. Rodríguez Rey, K. Mu, M. Garcia González, A. Bernárdez Braña, and X. Blanco Escoda. 2024a. iread4skills dataset 2: annotated corpora by level of complexity for fr, pt and sp.

A. Pintard, Thomas François, J. Nagant de Deuxchaisnes, Sílvia Barbosa, M. L. Reis, Michel Moutinho, Ricardo Monteiro, Raquel Amaro, Sara Correia, S. Rodríguez Rey, Marcos Garcia González, K. Mu, and Xavier Blanco Escoda. 2024b. iread4skills dataset 1: corpora by complexity level for fr, pt and sp.

E. Ribeiro, N. Mamede, and J. Baptista. 2024a. Automatic Text Readability Assessment in European Portuguese. In *PROPOR 2024*, pages 97–107.

E. Ribeiro, N. Mamede, and J. Baptista. 2024b. Text Readability Assessment in European Portuguese: A Comparison of Classification and Regression Approaches. In *PROPOR 2024*, pages 551–557.

Andrew Steeds. 2001. *Adult Literacy Core Curriculum Including Spoken Communication*. ERIC/The Basic Skills Agency, Commonwealth House, London.

R. Wilkens, A. Pintard, T. François, S. Barbosa, M. L. Reis, R. Amaro, E. Ribeiro, N. Mamede, J. Baptista, X. Blanco, A. Catena, R. Gauchola, and K. Mu. 2024. iread4skills - basic lexicons per complexity level.

[1] https://iread4skills.com/tools-resources

[2] https://zenodo.org/communities/iread4skills/

Lexicon-Grammar Web

Jorge Baptista[1,2], **David Antunes**[1], **Nuno Mamede**[1], **and Eugénio Ribeiro**[1,3]
[1] INESC-ID Lisboa, Portugal
[2] Faculdade de Ciências Humanas e Sociais, Universidade do Algarve, Portugal
[3] Instituto Universitário de Lisboa (ISCTE-IUL), Portugal
{jorge.baptista,david.f.l.antunes,nuno.mamede,eugenio.ribeiro}@inesc-id.pt

Abstract

This demo showcases a web-based interface that provides open, interactive access to a large-scale grammatical database of European Portuguese verbal constructions. Through a unified search and exploration environment, users can query, inspect, and compare more than 7,000 distributionally free verbal constructions and over 2,700 verbal idioms (frozen constructions), grounded in long-standing Lexicon–Grammar descriptions. For each construction, the interface exposes core linguistic properties such as argument structure, distributional constraints, semantic roles, major syntactic transformations, and curated usage examples with English translations. The demo illustrates how detailed, manually validated grammatical knowledge can be explored dynamically via the web, supporting linguistic research, language teaching, and NLP development. To the best of our knowledge, this is the largest publicly accessible, web-based grammatical resource dedicated to European Portuguese verbal constructions.

1 Overview: Web Interface and Resources

The resources presented in this demo originate from long-term research of European Portuguese within the Lexicon–Grammar framework (Gross, 1996), with a strong orientation toward computational processing. The description of distributionally free verbal constructions stems from the systematic development of a Lexicon–Grammar of EP verbs, leading to the creation of the VIPER database (Baptista, 2013). This work resulted in a large, manually curated inventory of verb constructions, later consolidated in the *Dicionário Gramatical de Verbos do Português Europeu* (Baptista and Mamede, 2020). The corresponding web-accessible resource is made available through the STRING NLP chain (Mamede et al., 2012)[1] and the PORTULAN/CLARIN research infrastructure[2].

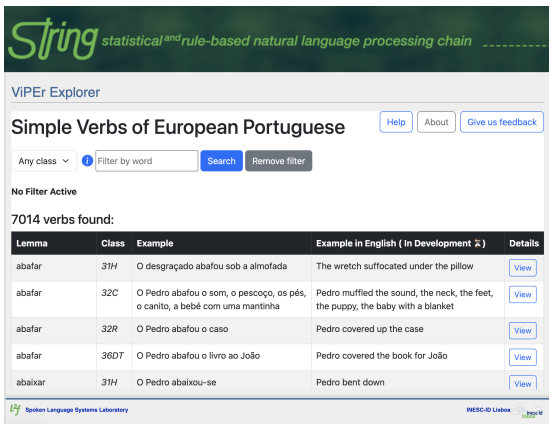
The VIPER resource encodes the +7,000 most frequently used EP verbal constructions. The information represented includes: (i) basic syntactic structure (e.g. impersonal, intransitive, transitive, ditransitive, and transitive–predicative patterns), covering more than 70 formal classes; (ii) explicit argument structure and distributional constraints, including human vs. non-human noun distinctions and over 70 semantic features corresponding to semantic prototypes (Bick, 2009); (iii) semantic roles selected from a set of approximately 50 highly reproducible constructs (Talhadas et al., 2013); (iv) constraints on subclause modality (indicative, subjunctive, factive, and *se*-interrogative) and coreference; and (v) major syntactic transformations, including several passive types, pronominalization and restructuring processes, symmetry constructions (Baptista, 2005), intrinsically reflexive constructions (e.g. *queixar-se* 'complain'), and *verba dicendi* constructions (Baptista, 2010). Each entry is illustrated with a canonical example and its English translation, complemented by manually curated corpus and web-based examples. Fig. 1 illustrates (a) the Search view of the VIPER and (b) the entry view of verb *agradar* 'please' (class 05).

The VIPER database has supported several NLP tools and linguistic studies integrated into the STRING system, including work on EP verb sense disambiguation (Pires, 2016), a survey on EP communication predicates (Reis et al., 2021), a study on transitive–predicative constructions (Baptista, 2021), and contrastive studies of locative verbs in European and Brazilian Portuguese (Rodrigues et al., 2015). It also serves as the catalogue of verbal senses for a lexicalized Abstract Meaning Representation initiative (Baptista et al., 2024). Together, these studies highlight the relevance of de-

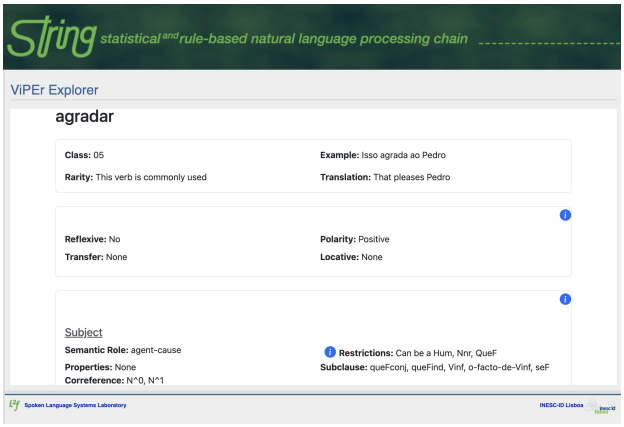
[1] https://string.hlt.inesc-id.pt/

[2] https://hdl.handle.net/21.11129/0000-000D-F91E-A

Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)
Vol. 2, pages 8–10
April 13-16, 2026



(a) Search view



(b) Entry view

Figure 1: VIPER web interface.

tailed lexical–syntactic descriptions for both theoretical investigation and applied NLP research, now made openly accessible through the web interface presented in this demo.

While grammatical descriptions of verbal constructions exist for both major varieties of Portuguese, many are available only in hard-copy form (Busse, 1994; Cançado et al., 2013) or as downloadable datasets primarily targeting Brazilian Portuguese, such as VERBNET.BR (Scarton, 2011) or VaLexPB[3]. In contrast, few resources for European Portuguese are manually curated, natively developed for EP, and provide comparably fine-grained linguistic descriptions in a web-accessible format.

In parallel, a second resource (VIDIOM) focusing on +2,700 verbal idioms has been developed. Verbal idioms are elementary sentences in which the verb and at least one of its arguments are distributionally frozen and the overall meaning is often non-compositional, that is, it is different from the meanings of the individual elements taken in isolation (e.g., *bater a bota*, lit. 'beat the boot', meaning 'to kick the bucket / die'). Fig. 2 shows the (a) the Search view, for a query with verb *bater* 'beat', and (b) the Entry view of the verbal idiom *bater a bota* 'die'. This work builds upon earlier, formal linguistic descriptions of frozen sentences in Portuguese, and has been progressively extended to support computational processing (Baptista et al., 2014). Recent developments include the integration of verbal idioms into rule-based parsers (Galvão et al., 2019), the creation of a corpus annotated for verbal idioms (Antunes et al., 2025b)[4], and dedicated methods for their automatic processing (Antunes et al., 2025a).

2 System Design

The system is implemented as a lightweight web application with a clear separation between data storage, backend logic, and user interface. The data layer relies on a SQLite relational database populated from structured spreadsheet sources and accessed through SQLAlchemy. The backend is implemented in Python using Flask, with modular routing via blueprints and server-side rendering through Jinja2 templates. The frontend consists of server-rendered HTML enhanced with Bootstrap components, enabling responsive layouts and interactive elements without reliance on a heavy client-side framework.

3 What the demo shows

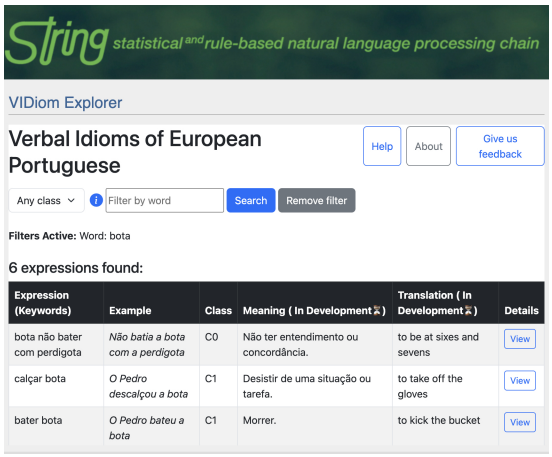
The demo provides interactive access to the Lexicon–Grammar resources via a unified web interface. Users can search and browse verbal constructions, examine their syntactic and semantic properties, and compare distributionally free constructions with verbal idioms, illustrated by curated examples and English translations to support analysis.

Acknowledgments

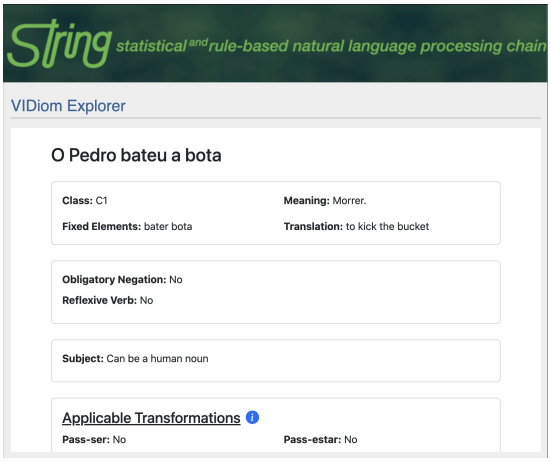
Work supported by national funds through Fundação para a Ciência e a Tecnologia, I.P. (FCT) (DOI: https://doi.org/10.54499/UID/50021/2025 and https://doi.org/10.54499/UID/PRR/50021/2025).

[3] https://github.com/jessemourao/VaLexPB/

[4] https://portulanclarin.net/repository/browse/vidiom-pt/



(a) Search view



(b) Entry view

Figure 2: VIDIOM web interface.

References

David Antunes, Jorge Baptista, and Nuno Mamede. 2025a. Processamento automático de expressões idiomáticas do português europeu. *Linguamática*, 17(1):55–73.

David Antunes, Jorge Baptista, and Nuno J. Mamede. 2025b. A European Portuguese corpus annotated for verbal idioms. In *Workshop on Multiword Expressions (MWE 2025)*, pages 58–66.

Jorge Baptista. 2005. Construções simétricas. In *Estudos de homenagem a Mário Vilela*, pages 353–367. FLUP.

Jorge Baptista. 2010. *Verba dicendi*: A structure looking for verbs. In *Les Tables. La grammaire du français par le menu. Mélanges en hommage à Christian Leclère*, pages 11–20. CENTAL/Presses Universitaires de Louvain.

Jorge Baptista. 2013. ViPEr: uma base de dados de construções léxico-sintáticas de verbos do Português europeu. In *Actas do XXVIII Encontro da APL - Textos Selecionados*, pages 111–129, Lisboa.

Jorge Baptista. 2021. Construções de verbos transitivos-predicativos em português: uma perspetiva transformacional. *Rev. da Assoc. Port. Linguística*, 8:10–25.

Jorge Baptista and Nuno Mamede. 2020. *Dicionário Gramatical de Verbos do Português Europeu*. U.Algarve Editora.

Jorge Baptista, Nuno Mamede, and Ilia Markov. 2014. Integrating verbal idioms into an NLP system. In *PROPOR 2014*, pages 250–255. Springer Verlag.

Jorge Baptista, Sónia Reis, João Dias, and Pedro Santos. 2024. Lexicalized Meaning Representation (LMR). In *5th Int. Workshop on Designing Meaning Representations @ LREC-COLING 2024*, pages 101–111.

Eckhard Bick. 2009. Semantic prototype tags for nouns. Online technical description.

Winfried Busse. 1994. *Dicionário Sintáctico de Verbos*. Almedina, Coimbra.

Márcia Cançado, L. Godoy, and L. Amaral. 2013. *Catálogo de Verbos do Português Brasileiro. Verbos de Mudança*. Editora UFMG.

Ana Galvão, Jorge Baptista, and Nuno J. Mamede. 2019. Processing European Portuguese Verbal Idioms: From the Lexicon-Grammar to a Rule-based Parser. In *3rd Int. Conf. Computational and Corpus-based Phraseology (EUROPHRAS 2019)*, pages 70–77.

Maurice Gross. 1996. Lexicon-grammar. In *Concise Encyclopedia of Syntactic Theories*, pages 244–259. Pergamon.

Nuno Mamede, Jorge Baptista, Cláudio Diniz, and Vera Cabarrão. 2012. STRING - A Hybrid Statistical and Rule-Based Natural Language Processing Chain for Portuguese. In *PROPOR 2012*, volume Demo Session, Coimbra, Portugal. PROPOR.

Ricardo Pires. 2016. Verb Sense Disambiguation in STRING. Master's thesis, Instituto Superior Técnico, Univ. Lisboa.

Sónia Reis, Nuno Mamede, and Jorge Baptista. 2021. Predicados de Comunicação em Português Europeu. *Rev. da APL*, 8:237–259.

Roana Rodrigues, Jorge Baptista, and Oto Araújo Vale. 2015. Análise contrastiva dos verbos locativos. In *Symposium in Information and Human Language Technology & IV Jornada de Descrição do Português*, pages 233–240.

Carolina Scarton. 2011. VerbNet.Br: construção de um léxico de verbos. In *Brazilian Symposium in Information and Human Language Technology*.

Rui Talhadas, Nuno Mamede, and Jorge Baptista. 2013. Semantic Roles for Portuguese Verbs. In *Int. Conf. on Lexis and Grammar*, pages 127–132.

Bruna: A Real-Time Multimodal Voice Agent with Hybrid Reasoning

Evandro Fonseca

Blip

evandro.fonseca@blip.ai

Abstract

This paper describes Bruna, a data-centric smart voice assistant powered by multiple Large Language Models designed to support Stilingue and Blip products. Our architecture provides an enriched conversational experience, delivering strategic insights in real-time.

1 Introduction

Currently, data analysts and decision-makers frequently face cognitive overload in their daily workflows. Although (Fonseca et al., 2024) propose a copilot capable of providing suggestions based on conversational context, optimizing the experience for customer service agents we identify a gap regarding the processing of dynamic, multimodal contexts. This limitation motivates us to explore novel approaches designed to deliver a more immersive analytical experience.

To address this challenge, we present Bruna, a voice assistant designed to streamline data interaction by processing audio and visual inputs simultaneously. Unlike text only systems, Bruna enables users to interact naturally with complex data representations.

Our architecture leverages the Model Context Protocol (MCP) (Hou et al., 2025) to dynamically connect the agent with external tools. To ensure low latency and fluid interaction, we utilize an asynchronous pipeline based on Server Sent Events (SSE) and the Azure OpenAI Realtime API[1]. This allows the agent to perceive visual content, such as charts on a screen and cross reference it with live market data on demand, providing immediate, hands-free strategic insights.

2 Architecture

Our architecture is designed to be modular and agnostic regarding the underlying Large Language Models (LLMs). In our implementation, we adopted a composite strategy that leverages the strengths of multiple models. For the conversational interface, we utilized Azure OpenAI Realtime GPT-4o to ensure low-latency verbal interaction. However, in our experiments, we observed that models optimized for real-time audio processing often lack efficient reasoning capabilities for complex analytical tasks. Therefore, seeking support from more robust text-based models is a fundamental part of our strategy. To address this, the system offloads complex data processing and reasoning tasks to Gemini 2.5 Flash-Lite(Comanici et al., 2025), ensuring that the agent remains both responsive and intellectually capable.

When developing applications that integrate real-time audio with external execution tools, interoperability is a major challenge. At the time of our agent's conception, we did not find foundation models with native support for the Model Context Protocol (MCP). Consequently, it was necessary to construct a custom MCP Adapter. This component acts as a bridge, standardizing the communication between the proprietary real-time API and any standard MCP server. This allows our agent to connect seamlessly with diverse external tools, such as Radar Stilingue or vector databases, regardless of the underlying model's native capabilities.

Recent work by Mileff (Mileff, 2025) addresses the latency challenge in voice agents by proposing a parallelized architecture that orchestrates WebSockets and multi-threaded Text-to-Speech (TTS) services. While Mileff's approach effectively minimizes the delay between text generation and audio synthesis through a segmented pipeline, our architecture differs by adopting a native multimodal stream approach. Instead of optimizing the serialization of text-to-audio, we integrate the tool execution layer directly into the context loop. In this way, Bruna does not merely read generated text faster; she suspends the audio stream to "think"

[1]available in https://learn.microsoft.com/pt-br/azure/ai-foundry/openai/realtime-audio-quickstart

Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)
Vol. 2, pages 11–13
April 13-16, 2026

(process data via Gemini/MCP) only when deep reasoning is required, resuming the conversation with validated insights.

Figure 1 shows the complete pipeline. When a session is initiated, the MCP Adapter registers the available tools based on the user's permissions. Throughout the interaction, the audio input is processed by the Realtime model. If a complex intent is detected (e.g., "Analyze this dashboard"), the adapter intercepts the request, routes the visual and textual context to Gemini Flash 2.5 for reasoning, and returns the synthesized insight to the audio model for verbal delivery. This hybrid approach ensures that the system maintains the fluidity of a real-time voice agent while possessing the analytical depth of a large-scale text model.

3 Interface and Use Cases

Typically, extracting actionable intelligence from social listening platforms requires navigating through complex dashboards, applying multiple filters, and manually interpreting high-dimensional charts. This process can be time-consuming and cognitively demanding for decision-makers who need immediate strategic answers. Considering this friction, the Voice Agent Controller interface was designed not merely as a replacement for visual dashboards, but as a conversational abstraction layer that simplifies access to complex data streams.

As illustrated[2] in Figure 2, the web interface is designed to be minimalist, prioritizing the audio-visual interaction stream. The layout consists of session controls (start/stop), media toggles (camera/microphone), and a granular real-time event log. This log serves a critical role in system transparency and explainability, displaying *Server-Sent Events* (SSE) regarding connection status, tool execution steps, and token consumption metrics. This allows the user to monitor the agent's "reasoning" process in real-time, validating that the correct external tools are being invoked before an audio response is synthesized.

3.1 Case Study: Strategic Marketing Analysis

To validate the effectiveness of the multimodal architecture and the reasoning capabilities of the hybrid model strategy, we conducted a case study

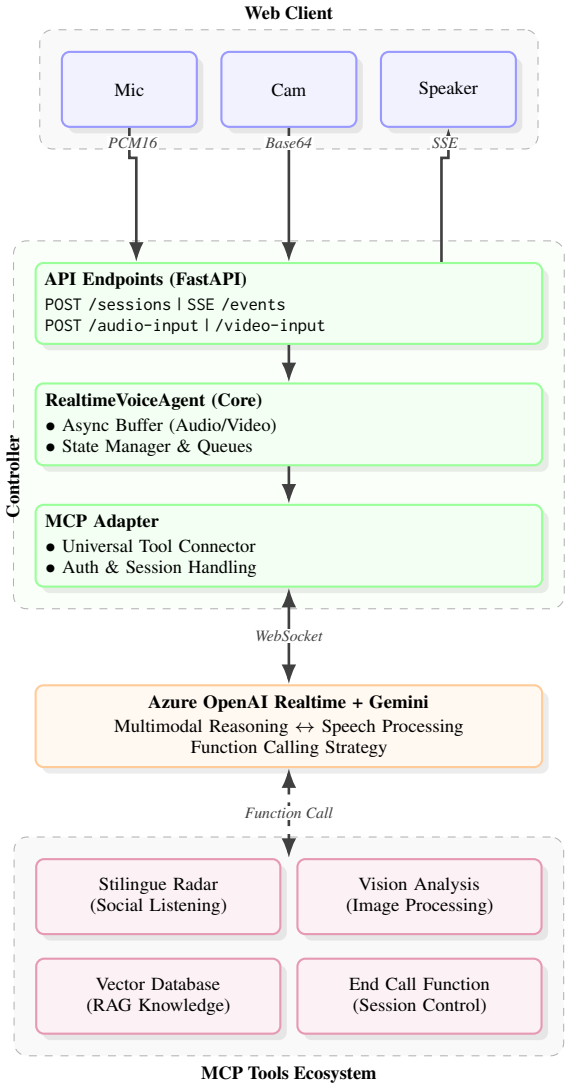


Figure 1: Architecture of the Voice Agent Controller. The system buffers multimodal inputs (audio/video) and orchestrates dynamic tool execution via the Model Context Protocol (MCP) Adapter. This layer facilitates Retrieval-Augmented Generation (RAG) (Lewis et al., 2020) and real-time data access to ground the model's reasoning before interacting with the Azure OpenAI Realtime API.

[2]Bruna's web interface is available to test in: https://secure-backend-api.stilingue.com.br/rd-envision-mcp-server/prod/voice-agent/test

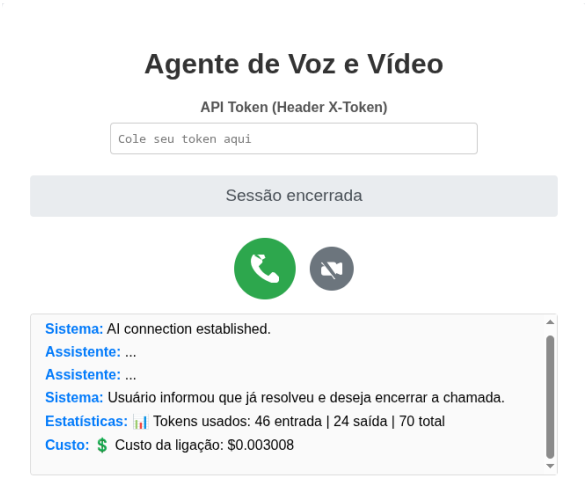


Figure 2: The Voice Agent Controller interface. The central panel manages the WebRTC media streams, while the bottom log provides real-time feedback on system events, tool execution, and token usage costs.

simulating a strategic planning session. In this scenario, the user acts as a marketing manager for a major fashion retail giant, aiming to structure a campaign for the Carnival holiday.

The interaction demonstrates the system's ability to convert raw data into strategic differentiation:

1. **Contextual Query:** The user requests a benchmark analysis of competitors in the apparel sector to guide a Carnival campaign strategy.

2. **Autonomous Execution:** The agent identifies the need for real-time data and triggers the social listening tool via MCP to query live sentiment and trending topics.

3. **Insight Generation:** The analysis detects a market opportunity: high positive engagement for "promotions" but significant negative sentiment regarding "delivery delays" among competitors.

4. **Strategic & Creative Output:** Bruna synthesizes a strategy focusing on reliability to counter competitor weaknesses and proposes campaign names like *"Carnaval sem Perrengue"* (Hassle-free Carnival).

This case study highlights the system's capacity to function as a "proactive copilot." By abstracting the complexity of database queries and sentiment analysis into a natural conversation, the interface allows users to focus on high-level decision-making rather than data mining.

4 Conclusion

In this paper, we presented Bruna, a multimodal voice agent that leverages the Model Context Protocol (MCP) and hybrid reasoning to generate real-time strategic insights. We detailed our architecture designed for low latency, showing how it reduces the cognitive load required to interpret complex data and maximizes the efficiency of decision-making processes. By maintaining conversational fluidity while accessing external tools, the system bridges the gap between static dashboards and active intelligence.

As further work, we intend to integrate the agent into the workflows of the Stilingue and Blip platforms. We also plan to conduct quantitative and qualitative studies to measure how the agent optimizes user time and reduces the "time-to-insight" compared to traditional visual dashboard navigation.

References

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Evandro B Fonseca, Tayane Soares, Dyovana Baptista, Rogers Damas, and Lucas Avanço. 2024. Blip copilot: a smart conversational assistant. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese-Vol. 2*, pages 194–196.

Xinyi Hou, Yanjie Zhao, Shenao Wang, and Haoyu Wang. 2025. Model context protocol (mcp): Landscape, security threats, and future research directions. *arXiv preprint arXiv:2503.23278*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Péter Mileff. 2025. Real-time, low audio latency based ai-powered application architecture design. *Production Systems and Information Engineering*, 13(1):46–63.

FlowDisco: Interactive Exploration of Dialogue Flows

Patrícia Ferreira[1,2], **Carolina Loureiro**[2], **Ana Alves**[1,3]
Catarina Silva[1,2], **Hugo Gonçalo Oliveira**[1,2]

[1]CISUC/LASI – Centre for Informatics and Systems of the University of Coimbra, Portugal
[2]Department of Informatics Engineering, University of Coimbra, Portugal
[3]Polytechnic Institute of Coimbra, Coimbra Institute of Engineering (ISEC), Portugal
{patriciaf, ana, catarina, hroliv}@dei.uc.pt
carolina.g.loureiro@gmail.com

Abstract

Analyzing large conversational datasets is often inefficient due to the linear nature of text, which hinders the tracking of interaction evolution over time. To address this, we present FlowDisco, an interactive platform for the automatic discovery and exploration of dialogue flows. The framework uses semantic embeddings and modular clustering to transform raw text into probabilistic dialogue flows. By providing a web interface with dynamic filtering and a suite of analytical metrics, FlowDisco simplifies the visual identification and validation of conversational behaviors at scale. The platform's utility is demonstrated through real-world application scenarios, including customer support interactions and multi-party political debates, where it successfully uncovers complex patterns and sentiment shifts that traditional sequential analysis often overlooks.

1 Introduction

The increasing volume of conversational data makes manual analysis inefficient and difficult to scale (Ammar and Bennani, 2025). Traditional NLP techniques often treat utterances as isolated events, hindering the tracking of interaction patterns across thousands of dialogues (Cardoso et al., 2025). Linear reading prevents an understanding of structural dynamics and the multiple paths a conversation can take, creating a need for tools that bridge the gap between a global flow view and the original text (Ganesh et al., 2023; Bouraoui et al., 2019).

To address this, we present FlowDisco, an interactive platform for the automatic discovery and exploration of dialogue flows. It transforms raw data into probabilistic transitions through a modular pipeline combining semantic vectorization with clustering algorithms (Ferreira et al., 2025). The framework provides an intuitive web interface for real-time manipulation of interaction networks, simplifying the identification of behaviors at scale (Bouraoui et al., 2019). By utilizing probability thresholds and node inspection, users can explore complex trajectories and validate model coherence through real utterance examples.

The practical utility of FlowDisco is demonstrated through two distinct scenarios: Study-AI, a dataset with task-oriented human-machine dialogues from a Portuguese schoolbook campaign and complex multi-party debates from the 2025 Portuguese Parliament. These scenarios showcase the platform's ability to handle diverse domains and dynamics, ranging from commercial customer support to adversarial political discourse.

2 FlowDisco

FlowDisco is a full-stack framework for the automatic discovery and interactive exploration of dialogue flows[1]. Its architecture consists of a backend for modular data processing and a frontend for visual analysis, transforming conversational data into structured probabilistic models. This setup facilitates the identification of underlying patterns, following the pipeline illustrated in Figure 1.

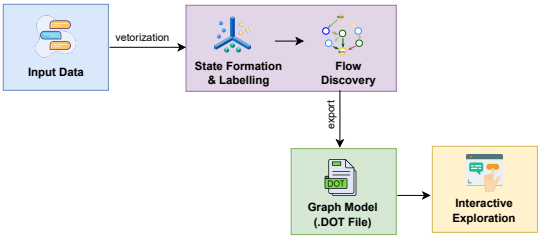


Figure 1: The FlowDisco processing pipeline.

2.1 Processing and Modeling Architecture

The backend, implemented in Python, transforms tabular datasets with three mandatory fields (dialogue_id, speaker, and utterance) into probabilistic graphs. First, utterances are

[1]Available at: https://github.com/NLP-CISUC/FlowDisco

Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)
Vol. 2, pages 14–17
April 13-16, 2026

vectorized using specific Sentence Transformers (e.g., a multilingual model like `paraphrase-multilingual-MiniLM-L12-v2`) to capture contextual meaning and accurately handle the specificities of the Portuguese language. These are then grouped into dialogue states ($s \in S$) through clustering algorithms like K-Means or DBSCAN, optimized by Silhouette Score or V-measure, or based on keywords and semantic categories. Once states are defined, they receive interpretable labels via strategies like LLMs, verb extraction, or KeyBERT [2]. The system then computes transition probabilities (p_{ij}) and dialogue-flow metrics for analysis. The final graph is exported as a .dot file for rendering. Additionally, FlowDisco can incorporate sentiment or metadata annotations to enrich the flow (Ferreira et al., 2024).

2.2 Interactive Frontend

The FlowDisco frontend, illustrated in Figure 2, provides a web-based dashboard for the interactive exploration of dialogue flows.

To manage the complexity of dense networks, the interface implements navigation controls and dynamic filtering. A central feature is the threshold slider, which allows for hiding transitions below a specific probability ($p_{ij} < \theta$), revealing the main conversation paths. Users can further isolate trajectories through speaker filters, by individual or group, or locate specific states via keyword node search. Navigation is supported by zoom, pan, and a selection-based box-zoom tool. For improved readability, nodes can be individually repositioned, and the original view can be restored at any time. Additionally, the upload panel allows for loading new .dot files, while the screenshot function saves the current visualization for external use.

The framework utilizes a visual encoding strategy where edge thickness is proportional to transition probability and a color gradient, from red to green, maps the sentiment flow. A legend facilitates the interpretation of these visual elements by identifying node types as either utterances or actions, marking special states such as the start and end of dialogues, and aiding in the identification of participants, which is particularly useful in multi-speaker datasets. To ensure model transparency, FlowDisco enables node inspection through contextual tooltips. Upon hovering, the system highlights incoming and outgoing paths and displays up to five real utterance examples, featuring a “Show More” button that indicates the remaining phrases available in the cluster. This allows for the validation of semantic coherence within groups without leaving the visualization.

A sidebar provides real-time statistical support through a metrics panel. This panel organizes information into Dialogue Metrics, including total statements, actions, and dialogues; Grouping Metrics, covering the number of dialogue and action states; and Flow Metrics, which track transition counts, sentiment data, and coverage metrics. These indicators are dynamically updated; high threshold values reduce visible nodes to focus on likely flows, while lower values increase graph coverage, allowing for the correlation of visual abstraction with the statistical precision of the model.

Overall, by combining automated modeling with interactive visualization, FlowDisco provides a comprehensive framework for understanding and analyzing conversational dynamics.

3 Study Cases

To validate the versatility of FlowDisco, we explored two cases with opposing characteristics to demonstrate how the interface enables extracting conclusions in different domains.

The first case, Study-AI, involves a customer support system with 105,376 utterances. The probability threshold slider was critical to hide rare transitions and focus on main paths. Sentiment colors allowed us to distinguish success flows (in green) from frustation zones (in red). Through node inspection, we found that red areas corresponded to topics where the system failed to solve problems or understand the user, creating negative repetitions that are difficult to detect through linear reading.

The second case, the Parliament, analyzes the 2025 State of the Nation debate, characterized by multi-speaker dynamics and long interactions. Speaker filters were fundamental to reducing visual noise and separating the polarities of each party. The tool also differentiates action nodes (such as Applause or Laughter) from speech nodes, identifying behavioral patterns like self-applause by specific party benches or systematic protests (e.g., consecutive interruption nodes like "False!" or "Shame!" clustered around specific political topics) that transcripts do not clearly evidence.

In both cases, the metrics panel validated vi-

[2] Maarten Grootendorst. KeyBERT: Minimal keyword extraction with BERT., 2020

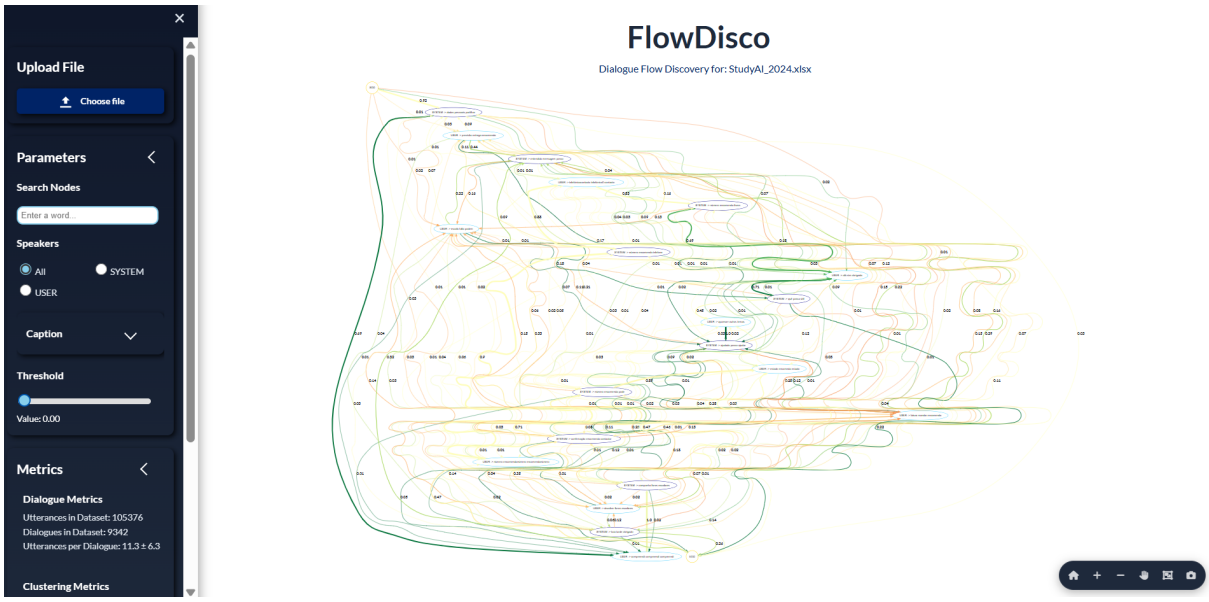


Figure 2: The FlowDisco interface, showcasing the sidebar with parameters and the main dialogue flow visualization.

sual observations. In Study-AI, a flow density of 0.33 confirmed a logical and closed structure despite the volume. In contrast, the low density in Parliament (0.12) corroborated the unpredictable nature of the debate. Monitoring node and transition counts allowed us to balance visual detail with statistical representativeness as we moved the threshold slider.

4 Conclusion

This work presented FlowDisco, an interactive tool designed to make the analysis of conversational data easier and more efficient. By transforming raw text into visual dialogue flows, the platform allows users to explore patterns that are usually hidden when reading text line by line. Through semantic clustering, probabilistic modeling, and dynamic filtering, FlowDisco provides a clear view of how conversations evolve, even in very large or complex datasets.

The application to two different datasets demonstrated the tool's flexibility. In the Study-AI case, the interface allowed for the rapid identification of points of dissatisfaction within thousands of short messages. In the Parliament case, the use of speaker filters and action nodes successfully organized the complex interactions of a multi-speaker debate. These results show that FlowDisco is a valuable framework for anyone needing to understand large-scale dialogue structures. Future work will focus on making the tool work with live data, allowing users to monitor and analyze conversations as they occur.

Acknowledgments

This work was financed by the Portuguese Recovery and Resilience Plan (PRR), through project C645008882-00000055 – Center for Responsible AI.

This work was also supported by FCT – Foundation for Science and Technology, I.P., within the scope of the research unit UID/00326 - Centre for Informatics and Systems of the University of Coimbra. Patrícia Ferreira was supported by FCT – Foundation for Science and Technology, I.P. through the PhD scholarship with reference 2024.01240.BD.

References

Mohamed Achref Ben Ammar and Mohamed Taha Bennani. 2025. A computational approach to modeling conversational systems: Analyzing large-scale quasi-patterned dialogue flows. In *IEEE EUROCON 2025-21st International Conference on Smart Technologies*, pages 1–6. IEEE.

Jean Léon Bouraoui, Sonia Le Meitour, Romain Carbou, Lina M Rojas Barahona, and Vincent Lemaire. 2019. Graph2bots, unsupervised assistance for designing chatbots. In *20th Annual SIGdial Meeting on Discourse and Dialogue*, pages 114–117. ACL.

Mauro Cardoso, Eugénio Ribeiro, and Fernando Batista. 2025. Portuguese far-right discourse on social media: Insights from topic modeling. In *14th Symposium on Languages, Applications and Technolo-*

gies (SLATE 2025), pages 12–1. Schloss Dagstuhl–Leibniz-Zentrum für Informatik.

Patrícia Ferreira, Isabel Carvalho, Ana Alves, Catarina Silva, and Hugo Gonçalo Oliveira. 2024. Sentiment-aware dialogue flow discovery for interpreting communication trends. In *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 274–288, Kyoto, Japan. Association for Computational Linguistics.

Patrícia Ferreira, Daniel Martins, Ana Alves, Catarina Silva, and Hugo Gonçalo Oliveira. 2025. Unsupervised flow discovery from task-oriented dialogues. In *Hybrid Intelligent Systems*, pages 336–347, Cham. Springer Nature Switzerland.

Ananya Ganesh, Martha Palmer, and Katharina von der Wense. 2023. A survey of challenges and methods in the computational modeling of multi-party dialog. In *Proceedings of the 5th Workshop on NLP for Conversational AI (NLP4ConvAI 2023)*, pages 140–154.

AttentionApp: An Interactive Tool for Analyzing Transformer Attention Patterns in Portuguese

Ricardo G. Oliveira[1], **Daniela Barreiro Claro**[1]

[1]FORMAS Research Center on Data and Natural Language
Institute of Computing – Federal University of Bahia (UFBA) – Salvador - Bahia - Brazil
{gomesricardo, dclaro}@ufba.br

Abstract

This paper presents *AttentionApp*, an interactive demonstration system designed to support the inspection and linguistic analysis of attention mechanisms in Transformer-based language models for Portuguese. The tool allows users to input sentences in Portuguese and visualize attention distributions across layers and heads, enabling fine-grained qualitative analysis of syntactic and semantic patterns captured by the model. AttentionApp is intended as a research-oriented tool, facilitating exploratory analysis, hypothesis generation, and interpretability studies for Portuguese Natural Language Processing.

1 Introduction

The introduction of the Transformer architecture based on self-attention mechanisms (Vaswani et al., 2017) have become a dominant paradigm in Natural Language Processing (NLP) domain. This paradigm has been successfully adopted in large pre-trained models such as BERT (Devlin et al., 2019), including language-specific variants for Portuguese, such as BERTimbau (Souza et al., 2020).

Despite their empirical success, understanding how these models encode linguistic structure remains an open research problem. Attention mechanisms, while not explanations per se, provide a useful lens for exploratory analysis and qualitative inspection of model behavior, as demonstrated in prior analyses of attention distributions in BERT-based models (Clark et al., 2019; Voita et al., 2019), including recent studies focusing on syntactic patterns in Portuguese (Oliveira et al., 2025).

For Portuguese, this challenge is amplified by linguistic phenomena such as rich verbal morphology, flexible word order, clitic placement, and complex agreement patterns. However, most interpretability tools and visualization frameworks have been developed with English-centric assumptions, limiting their applicability to Portuguese.

In this demonstration, we present *AttentionApp*, an interactive tool designed to visualize and analyze attention patterns produced by Transformer models when processing Portuguese text. The system aims to support researchers, students, and practitioners interested in interpretability, syntactic analysis, and information extraction for Portuguese.

2 System Overview

AttentionApp is implemented as a modular Python-based system with a web-oriented interactive interface. The architecture is composed of three main components:

- **Preprocessing and Tokenization Module**, responsible for sentence segmentation, subword tokenization, and token alignment, following the standard tokenization schemes used in BERT-style models (Devlin et al., 2019).
- **Model and Attention Extraction Module**, which loads Transformer-based language models for Portuguese, such as BERTimbau (Souza et al., 2020), and extracts attention weights across layers and heads.
- **Visualization and Interaction Module**, which renders attention distributions in an interactive format, enabling exploration of token-to-token relations.

The system was designed with extensibility in mind, allowing future integration of additional models, linguistic annotations, or downstream tasks.

3 Main Functionalities

AttentionApp provides the following core functionalities:

- Sentence-level input in Portuguese, allowing users to analyze arbitrary examples.

Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)
Vol. 2, pages 18–20

- Extraction of attention weights from all layers and attention heads of the selected Transformer model.
- Interactive visualization of attention matrices, supporting token-level inspection and comparison across heads, as commonly adopted in attention analysis studies (Clark et al., 2019).
- Layer and head selection, enabling fine-grained analysis of attention specialization, in line with findings on head-level functional differentiation (Voita et al., 2019).
- Exploratory linguistic analysis, allowing users to qualitatively assess whether attention patterns align with syntactic or semantic relations, following methodologies previously applied to syntactic analysis through attention heads (Oliveira et al., 2025).

These features make the tool suitable for exploratory research, pedagogical use, and qualitative evaluation of model behavior.

4 Demonstration Setup

During the demonstration session, participants will interact directly with AttentionApp through a live interface. Users will input Portuguese sentences and dynamically explore the resulting attention visualizations by selecting layers and heads.

The demonstration will focus on illustrating how different attention heads capture distinct relational patterns and how these patterns vary across layers, as previously observed in empirical analyses of Transformer attention (Clark et al., 2019; Voita et al., 2019). The system runs locally and does not require specialized hardware beyond a standard laptop, ensuring robustness in a conference environment.

5 System Availability

AttentionApp is an open-source research tool and its complete source code is publicly available on GitHub. The repository includes installation instructions, dependency specifications, and example configurations that allow the system to be executed locally.

The project is available at:

`https://github.com/rgoliveirati/attention_app`

This availability ensures transparency, reproducibility, and facilitates further extensions by the research community.

AttentionApp is an open-source research tool and its complete source code is publicly available on GitHub (Oliveira, 2025).

6 Linguistic Relevance and Applications

AttentionApp is particularly relevant for research on Portuguese NLP, as it enables direct inspection of model behavior on language-specific constructions processed by Transformer-based architectures (Vaswani et al., 2017; Devlin et al., 2019). Potential applications include:

- Analysis of syntactic dependencies and word order effects.
- Support for research on attention-based information extraction.
- Qualitative evaluation of model interpretability.
- Pedagogical use in computational linguistics and deep learning courses.

By focusing explicitly on Portuguese and leveraging language-specific pre-trained models (Souza et al., 2020), the tool contributes to reducing the methodological gap between high-resource and less-resourced languages in interpretability research.

7 Conclusion

This demonstration presents *AttentionApp*, an interactive tool for analyzing attention mechanisms in Transformer models applied to Portuguese. By enabling direct and fine-grained inspection of attention patterns, the system supports exploratory linguistic analysis and interpretability-oriented research. AttentionApp is intended as a practical resource for the Portuguese NLP community and aligns with the goals of the PROPOR Demonstration Track.

8 Acknowledgments

This work was supported by FAPESB (TIC002/2015 and CCE022/2023), CAPES, CNPQ, and INCT-TILDIAR/CNPq (408490/2024-1).

References

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What does BERT look at? an analysis of BERT's attention. In *Proceedings of the BlackboxNLP Workshop at EMNLP 2019*, pages 276–286. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2019)*, pages 4171–4186, Minneapolis, MN, USA. Association for Computational Linguistics.

Ricardo G. Oliveira. 2025. Attention app: Interactive visualization and analysis of attention heads in transformer models. `https://github.com/rgoliveirati/attention_app`. Source code repository.

Ricardo G. Oliveira, Daniela B. Claro, and Rerisson Cavalcante. 2025. Syntactic analysis in transformers through attention heads. In *Anais do XVI Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana (STIL 2025)*, pages 295–306, Fortaleza, CE, Brazil. Sociedade Brasileira de Computação.

Fabio Souza, Rodrigo F. Nogueira, and Roberto A. Lotufo. 2020. BERTimbau: Pretrained BERT models for brazilian portuguese. In *Intelligent Systems: 9th Brazilian Conference on Intelligent Systems (BRACIS 2020), Part I*, volume 12066 of *Lecture Notes in Computer Science*, pages 403–417, Rio Grande, Brazil. Springer.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pages 5998–6008. Curran Associates, Inc.

Elena Voita, Jean Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. 2019. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Proceedings of ACL 2019*, pages 5797–5808.

Sistema Multimodal de Apoio ao Gerenciamento de Riscos de Desastres

Hosana Iasmin Castro dos Santos Lucena[1], Gabriel Rocha dos Santos[1], Jady Lima da Silva[1], Ricardo José Matos de Carvalho[2], Patrick Terrematte[1],

[1] Instituto Metrópole Digital (IMD), Universidade Federal do Rio Grande do Norte (UFRN)
Natal-RN, Brasil
[2] Departamento de Engenharia de Produção,
Programa de Pós-graduação em Engenharia de Produção (PEP),
Grupo de Extensão e Pesquisa em Ergonomia (GREPE),
Núcleo Interdisciplinar de Pesquisas sobre Desastres (NUPED)
UFRN, Natal-RN, Brasil

Correspondence: patrick.terrematte@ufrn.br

Resumo

Este artigo tem como objetivo apresentar o sistema multimodal computacional, denominado de NOAH, para apoiar o gerenciamento de riscos de desastres (GRD) nas cidades brasileiras, considerando a necessidade de troca de informações e de comunicações estabelecidas entre os agentes públicos de GRD e os membros da população em situações de riscos e desastres. Este sistema está sendo desenvolvido através da aplicação da inteligência artificial (IA), integrando o chatbot ao processamento de linguagem natural (PLN), reconhecimento de fala, classificação de imagens e recuperação de informações por geração aumentada de recuperação(RAG). O sistema tem como foco a comunicação direta com a população via WhatsApp, permitindo a coleta de relatos em língua portuguesa nos formatos de texto, áudio e imagem. A contribuição prática do NOAH consiste na combinação de uma técnica de modelagem de tópicos(BERTopic) para classificação textual, Whisper Small para transcrição de áudio e redes neurais convolucionais Resnet50 para análise visual do tipo de incidente. Essa abordagem viabiliza o desenvolvimento de ferramenta prática e escalável para o apoio à tomada de decisão dos órgãos municipais de Proteção e Defesa Civil, que são responsáveis pelo GRD, contribuindo para uma resposta mais eficiente a situações de emergência em localidades de língua portuguesa.

1 Introdução

As mudanças climáticas globais estão tornando mais complexa a previsibilidade do comportamento climático e têm impactado socioeconomicamente a população em geral e, em especial, as populações mais vulneráveis aos desastres, que têm ocorrido com maior frequência e intensidade no planeta (United Nations, 2023).

O desastre é um fenômeno social resultante das ações dos seres humanos e de suas sociedades (Helsloot, 2006), constituindo-se um dos grandes problemas urbanos, pois produz impactos sociais, econômicos e ambientais no mundo todo e, especialmente, nas localidades onde vivem e trabalham os munícipes.

De acordo com o relatório da United Nations Office for Disaster Risk Reduction (UNDRR, 2025), indivíduos nascidos em 1990 possuem cerca de 63% de probabilidade de vivenciar ao menos uma inundação catastrófica ao longo da vida, enquanto aqueles nascidos em 2025 apresentam uma probabilidade ainda maior, estimada em aproximadamente 86%. Segundo a agência oficial de notícias das Nações Unidas, o custo total associado aos eventos climáticos extremos ultrapassa US$ 2,3 trilhões por ano, representando um impacto econômico cerca de dez vezes maior do que estimativas anteriores (UN News, 2025).

Recentemente, o governo brasileiro tem apresentado políticas públicas com investimentos financeiros para custear políticas preventivas de desastres, iniciadas pela Lei 12.608/2012 (Brasil, 2012), notadamente, que foram acentuadas, principalmente em 2024, 2025 e 2026, com a disponibilização de investimentos para a elaboração de Planos Municipais de Redução de Riscos e Desastres (PMRR) e de Planos de Adaptação Climática nos municípios brasileiros. Esse cenário torna ainda mais relevante o desenvolvimento de ferramentas tecnológicas capazes de apoiar a identificação precoce de situações de risco por parte das autoridades públicas, que podem causar o desastre, e, ainda, agilizar respostas emergenciais ou contingenciais mais eficientes e eficazes de órgãos municipais responsáveis pelo GRD, quais sejam, os órgãos municipais de proteção e defesa civil (OMPDC).

Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)
Vol. 2, pages 21–24

Atualmente, o registro de ocorrências de ameaças ou desastres pelo órgão de Proteção e Defesa Civil de Natal-RN é realizado de forma manual, inicialmente em formulário de papel e, depois, organizados em planilhas eletrônicas (excel), conforme observado durante coleta de informações para o sistema NOAH. Esse processo torna a classificação dos incidentes ou desastre, dependente unicamente de interpretação humana, o que pode introduzir inconsistências na categorização das ocorrências, além disso, a consolidação manual dos dados reduz a capacidade de realizar análises estatísticas ou espaciais automatizadas, em situações críticas essas limitações podem impactar negaticamente a priorização de ocorrências e alocação de recursos.

Diante desse cenário, este artigo apresenta o NOAH, um sistema colaborativo computacional de apoio ao gerenciamento de riscos de desastres, visando facilitarndo a comunicação entre agentes públicos e a população e a coordenação de ações pelos agentes públicos, como também otimizarndo a classificação do tipo de desastre por meio de mensagens escritas,em áudio ou imagens sobre riscos de desastres,

2 Trabalhos relacionados

Wibowo et al. (2025), em sua revisão bibliométrica, afirmam que a China é o país mais produtivo em aplicação de Inteligência artificial (IA) no gerenciamento de desastres e os Estados Unidos são os mais citados, e que foram identificados nesta revisão seis grupos de pesquisa, quais sejam: monitoramento e previsão de desastres usando redes IoT; tecnologia geoespacial baseada em IA para gestão de riscos; sistemas de apoio à decisão para gestão de emergências em desastres; análise de redes sociais para resposta a emergências; algoritmos de aprendizado de máquina para redução do risco de desastres; big data e aprendizagem profunda para gestão de desastres.

Sistemas digitais também têm sido propostos para facilitar a comunicação entre a população e as autoridades durante situações de desastre. Por exemplo, Nik Nazli et al. (2016) propuseram uma aplicação móvel que permite aos cidadãos registrar e informar ocorrências de desastres diretamente às autoridades responsáveis. O sistema possibilita o envio de descrições textuais, imagens e informações de localização geográfica obtidas por meio do GPS do dispositivo móvel, permitindo que gestores de emergência visualizem e acompanhem as ocorrências reportadas. Entretanto, o sistema baseia-se principalmente no registro manual das informações e não incorpora técnicas de inteligência artificial para o processamento automático dos dados enviados pelos usuários.

Uma análise recente de literatura sobre o uso de Processamento de Linguagem Natural para desastres demonstra que, apesar do crescente interesse, a aplicação de técnicas de mineração de texto, classificação e extração de dados permanece subutilizada em muitos contextos (Godinho, 2024). A revisão aponta que, embora métodos como classificação supervisionada, extração de entidades e modelagem de tópicos tenham sido aplicados com sucesso em alguns estudos, há limitações significativas, como escassez de dados anotados, dificuldades com linguagem informal e falta de representatividade regional.

Em paralelo, conjuntos de dados como o HumAID, criado pelo CrisisNLP(Firoj Alam, 2021), mostram que é possível coletar e anotar milhares de tweets de desastres, distribuídos em categorias humanitárias como “infraestrutura danificada”, “feridos”, “necessidades urgentes”, entre outras, permitindo o desenvolvimento de modelos automatizados de triagem e priorização. No entanto, esses recursos costumam estar restritos ao inglês ou a contextos internacionais, reforçando a lacuna para o português brasileiro, especialmente suas variedades regionais, justamente o espaço que este sistema busca explorar.

3 Arquitetura do Sistema NOAH

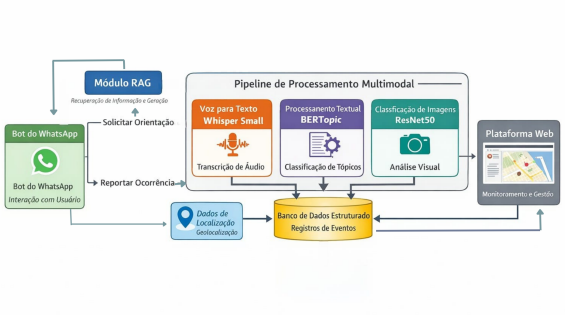


Figura 1: Arquitetura do Sistema NOAH com Módulo RAG, Bot do WhatsApp, pipeline de Processamento Multimodal e plataforma Web.

A interação do usuário com o sistema NOAH ocorre exclusivamente via WhatsApp, onde um bot conversacional atua como interface principal. Ao iniciar a comunicação, o usuário recebe duas

opções: (i) relatar uma ocorrência ou (ii) solicitar informações de orientação baseadas em planos oficiais de contingência. No segundo caso, o sistema utiliza uma abordagem (RAG) para recuperar e apresentar informações relevantes de acordo com o tipo de evento e o contexto informado.

Quando o usuário opta por relatar uma ocorrência, o sistema inicia um pipeline de processamento multimodal. Mensagens de texto são enviadas diretamente para a API de processamento de linguagem natural, enquanto mensagens de áudio e imagens seguem fluxos específicos de processamento. No caso de mensagens de áudio, o NOAH utiliza o modelo Whisper Small, desenvolvido pela OpenAI, para realizar a transcrição automática da fala para texto em língua portuguesa. Para aumentar a confiabilidade do processo, a transcrição gerada é apresentada ao usuário para confirmação antes de prosseguir. Após a validação, o texto transcrito é encaminhado ao módulo de classificação textual, garantindo que erros de reconhecimento de fala não comprometam a categorização da ocorrência.

As mensagens textuais, sejam elas enviadas diretamente pelo usuário ou provenientes da transcrição de áudio, são processadas pelo BERTopic (Grootendorst, 2022), ajustado para identificar categorias relacionadas a desastres e situações de risco. O modelo é responsável por atribuir uma classe semântica ao relato, padronizando descrições que, na linguagem natural, podem variar significativamente entre diferentes usuários.

Além de texto e áudio, o sistema também aceita imagens como forma de relato. As imagens enviadas são analisadas por uma rede neural convolucional baseada na arquitetura ResNet50, treinada para classificar cenários visuais associados a eventos críticos, como alagamentos, incêndios ou deslizamentos. Essa abordagem multimodal amplia o escopo do sistema, permitindo que evidências visuais complementem as informações textuais.

Após a classificação do conteúdo, o NOAH solicita ao usuário informações de localização, que podem ser compartilhadas diretamente pelo WhatsApp. Em seguida, todos os dados coletados — modalidade de entrada, conteúdo processado e localização — são armazenados como um registro estruturado no banco de dados do sistema. Esses registros alimentam uma plataforma web voltada à gestão e visualização das ocorrências, oferecendo suporte às ações da Defesa Civil.

4 Demonstração do sistema NOAH

A demonstração do NOAH foi concebida para simular cenários reais de comunicação entre cidadãos e órgãos de gerenciamento de riscos de desastres, especialmente o órgão municipal de proteção e defesa civil, além de SAMU, Corpo de Bombeiros, entre outros. Durante a apresentação, os participantes interagem com o sistema utilizando seus próprios dispositivos móveis, enviando mensagens ao número de WhatsApp associado ao bot, como fariam diante de uma situação real de perigo ou desastre.

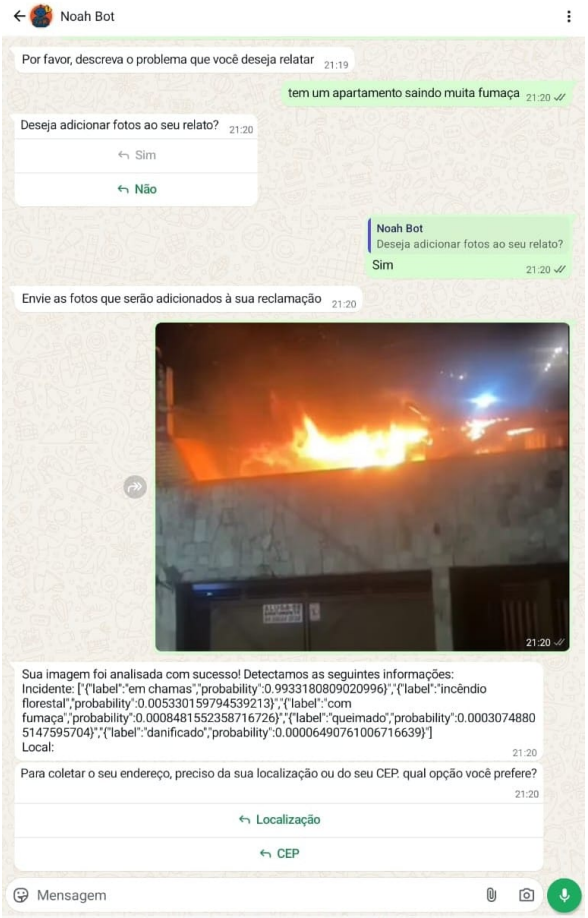


Figura 2: Exemplo de interação do usuário com o Noah Bot no WhatsApp para registro de um incidente. Após a descrição textual e envio de imagem pelo usuário, o sistema realiza a classificação automática da cena (detecção de incêndio) e solicita informações de localização para complementar o registro da ocorrência.

O fluxo da demonstração inicia-se com o envio de uma mensagem ao bot, que retorna automaticamente as opções de interação disponíveis. Ao selecionar a opção de relato de ocorrência, o participante pode submeter informações em diferentes formatos, incluindo texto, áudio ou imagem. Cada

tipo de entrada aciona o respectivo módulo de processamento, permitindo demonstrar a integração entre reconhecimento de fala, processamento de linguagem natural e classificação de imagens.

Durante a demonstração, mensagens de áudio são transcritas e apresentadas ao usuário para confirmação, evidenciando a preocupação do sistema com a confiabilidade das informações. Relatos textuais são classificados automaticamente, e imagens enviadas são analisadas pelo classificador visual. Em todos os casos, o sistema solicita a localização do evento antes de concluir o registro.

Por fim, os registros gerados são exibidos na interface web de gestão, onde é possível visualizar as ocorrências. A demonstração destaca a aplicabilidade prática do NOAH, bem como o papel central do processamento computacional do português na organização e interpretação de dados gerados pela população em contextos de emergência.

5 Conclusão

O NOAH fornece uma arquitetura modular e extensível que pode ser adaptada a diferentes cenários de uso, tanto para aplicações operacionais quanto para tarefas de anotação e análise de dados. Além de sua aplicação prática, o sistema possibilita a construção de um corpus em português, composto por interações reais entre usuários e o bot, o que representa uma contribuição relevante para pesquisas futuras em processamento de linguagem natural e sistemas no contexto de desastres.

Esperamos que este trabalho sirva como base para o desenvolvimento e a avaliação de sistemas interativos voltados à coleta de dados em ambientes críticos, contribuindo para uma melhor compreensão e evolução de soluções de PLN em português à medida que esses sistemas se tornam mais complexos e integrados ao mundo real.

Limitações

Uma limitação deste estudo refere-se à ausência, no estágio atual do projeto, de métricas quantitativas de avaliação do desempenho dos modelos propostos. Como o projeto ainda se encontra em desenvolvimento, não foi possível realizar uma validação empírica completa, principalmente devido à indisponibilidade de um conjunto de dados estruturado contendo registros detalhados de ocorrências que permitam testar e comparar sistematicamente os modelos. Dessa forma, a construção ou consolidação de um dataset adequado constitui um passo essencial para etapas futuras da pesquisa. Como trabalho futuro, pretende-se coletar, organizar e curar dados de ocorrências reais, possibilitando a aplicação de métricas de avaliação apropriadas e a validação rigorosa do desempenho dos modelos desenvolvidos.

References

Muhammad Imran Ferda Ofli Firoj Alam, Umair Qazi. 2021. Humaid: Human-annotated disaster incidents data from twitter. In *15th International Conference on Web and Social Media (ICWSM)*.

Matilde Martins Lencastre Godinho. 2024. The impact of natural language processing in disaster management: A systematic literature review. Master's thesis, Universidade NOVA de Lisboa (Portugal).

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *Preprint*, arXiv:2203.05794.

I. Helsloot. 2006. [review of: R.w. perry, e.l. quarantelli (2005) what is a disaster? new answers tot old questions]. *Journal of Contingencies and Crisis Management*, 14(1):55–56.

Nik Nadian Nisa Nik Nazli, Sapora Sipon, and Norita Md Norwawi. 2016. A prototype mobile application for informing disaster complaint – "informer on site". *International Journal of Interactive Mobile Technologies (iJIM)*, 10(1):68–70.

UN News. 2025. Relatório da onu alerta para aumento de eventos climáticos extremos. Acesso em: 15 mar. 2026.

UNDRR. 2025. Global assessment report on disaster risk reduction 2025: Resilience pays – investing and financing for our future. Accessed: 2026-03-15.

United Nations. 2023. Relatório anual das nações unidas. Acesso em: 15 mar. 2026.

Arief Wibowo, Ikhwan Amri, Asep Surahmat, and Rusdah Rusdah. 2025. Leveraging artificial intelligence in disaster management: A comprehensive bibliometric review. *Jàmbá: Journal of Disaster Risk Studies*, 17(1):a1776.

Lispector: Fine-tuning de Modelos de Linguagem para Revisão Gramatical e Ortográfica em Português Brasileiro

Andresa Medeiros[1], Felipe Iszlaji[1], Claudia Sarmiento-Moreno[1], Camila Muniz[1], Larissa Ponciano[1], Larissa Dejigov[1], Ronald Monteiro[1], Pedro Kretikouski[1], Guilherme Chaves[1]

[1]Clarice.ai

andresa.medeiros@clarice.ai, felipe@clarice.ai, clasarmor@gmail.com,
camilaamunizz@gmail.com, larissasponciano@gmail.com, contatolarissadc@gmail.com,
ronaldmonteiro.pro@gmail.com, pedro.junior@clarice.ai, guilherme.chaves@clarice.ai

Abstract

Este trabalho apresenta o Lispector, uma família de modelos de linguagem especializados para revisão gramatical e ortográfica em português brasileiro. Comparamos duas estratégias de inferência para a tarefa de correção gramatical de texto com grandes modelos de linguagem (LLMs): (1) *fine-tuning* supervisionado e (2) *prompting few-shot* em modelos de maior escala. Utilizando um conjunto de dados de 4.500 pares de textos reais de usuários (2.500 registros para treino, 1.000 para avaliação e 1.000 para teste), com referências corrigidas por linguistas, analisamos duas variantes do Lispector baseadas em diferentes tamanhos de parâmetros. A avaliação empregou as métricas BLEU, GLEU, METEOR e ROUGE. Os resultados demonstram que modelos menores submetidos a *fine-tuning* supervisionado superam consistentemente em todas as métricas modelos maiores que operam apenas com *prompting*, com o Lispector *small* alcançando ganhos expressivos em métricas de similaridade textual como GLEU (+12%) e BLEU (+13%). Assim, além do aumento de desempenho, os modelos *fine-tuned* apresentam comportamento mais previsível e conservador, características desejáveis em aplicações industriais de escrita assistida. No quesito latência, o Lispector *small* obteve a menor mediana de tempo de resposta entre todos os modelos e o menor P95 entre os *fine-tuned*; o Lispector *large* também se mostrou competitivo. Esses achados indicam que, para tarefas específicas de revisão textual em português brasileiro, o *fine-tuning* pode oferecer vantagens significativas em desempenho e eficiência computacional.

1 Introdução

A correção gramatical e ortográfica automática é uma tarefa fundamental em processamento de linguagem natural (PLN), com aplicações diretas em editores de texto e plataformas educacionais. Em contextos reais, sistemas de revisão textual devem identificar e corrigir erros de modo previsível, conservador e alinhado às normas linguísticas, evitando reformulações desnecessárias que comprometam a experiência do usuário. Trabalhos pioneiros em correção gramatical para português brasileiro utilizaram abordagens baseadas em regras e métodos híbridos, como ReGra (Martins et al., 1998), CoGrOO (Kinoshita et al., 2006) e LanguageTool (Naber, 2003). Com o advento dos grandes modelos de linguagem (LLMs), adotaram-se duas abordagens principais para correção textual: 1) *prompting few-shot* com modelos generalistas como GPT (Coyne et al., 2023; Fang et al., 2023), que não requer treinamento adicional, sendo atraente do ponto de vista industrial por sua facilidade de integração e baixo custo inicial; e (2) *fine-tuning* supervisionado, que adapta pesos de modelos pré-treinados para correção gramatical (Bryant et al., 2023; Rothe et al., 2022), que, embora exija maior investimento em dados e treinamento, permite maior especialização. Embora modelos maiores via *prompting* sejam frequentemente considerados superiores, com viabilidade para português brasileiro (Penteado and Perez, 2023), há evidências limitadas sobre a eficácia comparativa entre *prompting* e *fine-tuning* para correção textual nessa língua. Permanece em aberto até que ponto modelos menores com *fine-tuning* supervisionado podem competir com modelos maiores usando apenas *prompting*. Essa lacuna é relevante para aplicações industriais, com restrições de custo computacional e latência.

2 Metodologia

2.1 Dados

Os dados foram obtidos de textos reais submetidos por usuários na plataforma de edição de texto Clarice.ai. Cada instância consiste em um par formado por texto original e versão corrigida, pro-

Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)
Vol. 2, pages 25–29

duzida por linguistas para eliminar erros de ortografia, gramática e pontuação seguindo as normas do português brasileiro e preservando o conteúdo semântico original. O conjunto de dados contém 4.500 registros, divididos em 2.500 para treinamento, 1.000 para avaliação e 1.000 para teste, sem sobreposição entre conjuntos.

2.2 Modelos

Os experimentos compararam dois grupos principais de modelos: (i) a família Lispector, submetida a *fine-tuning* supervisionado para revisão gramatical e ortográfica em português brasileiro, e (ii) modelos de grande escala utilizados como *baselines*, exclusivamente via *prompting few-shot*, sem ajuste adicional de pesos. Importa notar que os modelos, dados e código não estão disponíveis publicamente por envolverem informações proprietárias e confidenciais vinculadas à Clarice.ai. A família Lispector inclui as variantes Lispector *large* (baseada no GPT-4.1), de maior porte, e Lispector *small* (baseada no GPT-4.1 nano). A segunda investiga se um modelo significativamente menor com *fine-tuning* supervisionado poderia alcançar desempenho comparável ou superior ao de modelos maiores que operam apenas via *prompting*. Essa comparação é relevante no contexto industrial, em que restrições de custo, latência e escalabilidade favorecem modelos mais compactos. Como *baselines*, utilizamos os modelos GPT-5, GPT-4.1, GPT-5 nano e GPT-4.1 nano, todos sem *fine-tuning* e avaliados somente via *prompting few-shot*, sendo que os mesmos prompts são usados para todos os modelos de *baseline*. Por razões de propriedade intelectual, os *prompts* específicos utilizados não podem ser divulgados. Em linhas gerais, todos os modelos *baseline* receberam *prompts* estruturados com instruções explícitas para correção gramatical e ortográfica em português brasileiro, incluindo exemplos representativos e orientações para preservar o conteúdo semântico original e evitar reformulações desnecessárias. Todos os modelos foram avaliados sobre o mesmo conjunto de dados de teste, permitindo comparação direta entre *fine-tuning* supervisionado e uso direto de modelos generalistas, e análise do *trade-off* entre capacidade do modelo, especialização para a tarefa e viabilidade industrial.

2.3 Configuração de treinamento

A Tabela 1 apresenta os hiperparâmetros do *fine-tuning* das duas variantes do Lispector. Buscou-se constância de parâmetros entre os modelos, de forma a permitir uma comparação controlada.

2.4 Métricas de avaliação

Os modelos foram avaliados por métricas automáticas escolhidas para contemplar a literatura consolidada de correção gramatical (GEC) e atender a requisitos de aplicações comerciais de edição de texto, como previsibilidade e consistência, evitando reformulações desnecessárias. Utilizamos BLEU (Papineni et al., 2002) e GLEU (Napoles et al., 2015) para medir a precisão de n-gramas, comparando as saídas dos modelos e os textos de referência. O GLEU é particularmente relevante por penalizar alterações desnecessárias, considerando, em simultâneo, a correção de erros e a preservação de trechos corretos. METEOR (Banerjee and Lavie, 2005) foi incluído por considerar correspondências parciais, enquanto ROUGE-1 e ROUGE-2 medem a sobreposição de unigramas e bigramas, respectivamente, permitindo uma análise complementar da preservação estrutural. Adicionalmente, avaliamos a latência de inferência de cada modelo, reportando média, mediana, P95 e valores mínimo e máximo de tempo de resposta em milissegundos. Essas métricas complementam a avaliação de qualidade textual com uma perspectiva de viabilidade operacional, relevante para aplicações industriais com restrições de tempo de resposta. Em conjunto, essas métricas equilibram rigor formal e tolerância a variações linguísticas naturais, importante no português brasileiro, caracterizado por múltiplas formas corretas de realização textual.

3 Resultados

3.1 Comparação geral

A Tabela 2 apresenta os resultados comparativos entre os modelos *baseline* (*prompting few-shot*) e os com *fine-tuning* supervisionado. Observa-se que os modelos da família Lispector superaram de modo consistente todos os *baselines*, indicando maior proximidade lexical e estrutural às correções de referência. Em BLEU, os Lispectors alcançaram 83% contra 70-77% dos GPTs; em GLEU, 89% contra 77-81%. Os modelos Lispector também obtiveram valores superiores em METEOR e ROUGE. Em particular, o desempenho do Lispector *small* é comparável ao do Lispector *large* em quase todas as métricas, apesar do menor número de parâmetros.

A Tabela 3 apresenta os resultados de latência. O

Parâmetro	**Lispector *large***	**Lispector *small***
Modelo base	GPT-4.1	GPT-4.1 nano
Épocas	3	3
Batch size	6	6
Learning rate mult.	0.1	0.1
Tokens treinados	—	~792.000

Table 1: Hiperparâmetros de treinamento.

Modelo	**BLEU**	**GLEU**	**METEOR**	**R-1**	**R-2**
Fine-tuned (zero-shot)					
Lispector *large*	83%	89%	97%	97%	90%
Lispector *small*	83%	89%	97%	97%	89%
Prompting (few shot)					
GPT-5	71%	78%	95%	94%	83%
GPT-4.1	70%	77%	96%	95%	81%
GPT-5 nano	74%	79%	93%	94%	85%
GPT-4.1 nano	77%	81%	94%	94%	87%

Table 2: Resultados comparativos entre modelos fine-tuned e prompting.

Lispector *small* (ft-lispector-small) obteve a menor mediana de tempo de resposta entre todos os modelos (1.378 ms) e o menor P95 entre os modelos *fine-tuned* (2.601 ms), superando inclusive modelos sem *fine-tuning* de porte equivalente. O Lispector *large* (ft-lispector-large) também se mostrou competitivo (mediana de 2.040 ms). Em contraste, os modelos da família GPT-5 apresentaram latências significativamente maiores, com medianas entre aproximadamente 20 e 41 segundos e P95 entre 33 e 162 segundos — valores incompatíveis com uso em produção em editores de texto.

3.2 Análise

A comparação entre *fine-tuning* supervisionado e *prompting few-shot* revela diferenças claras no comportamento dos modelos. Os modelos Lispector, com *fine-tuning* supervisionado, superaram consistentemente todos os modelos operados via *prompting*, com diferenças mais acentuadas em métricas de precisão lexical (BLEU: +13%, GLEU: +12%). Isso sugere que os modelos *fine-tuned* preservam melhor a forma das correções de referência, o que é desejável em aplicações de revisão textual, nas quais alterações desnecessárias compromem a confiança do usuário. Como achado relevante, o Lispector *small*, baseado em um menor tamanho de parâmetros (GPT-4.1 nano), alcançou desempenho equivalente ao Lispector *large* (baseado no GPT-4.1 completo) e superou de forma significativa modelos maiores que operam via *prompting*, em métricas de qualidade e de latência. Em termos de estabilidade, o Lispector *small* apresentou P95 de 2.601 ms, indicando comportamento consistente mesmo nos piores casos — contraste direto com os modelos GPT-5, cujo P95 chega a 162 segundos. Esse conjunto de resultados indica que, para tarefas específicas de revisão textual, a adaptação supervisionada pode ser mais eficaz que o uso direto de modelos generalistas de grande porte, tanto em qualidade quanto em viabilidade operacional.

4 Discussão

4.1 Implicações para aplicações industriais

Os resultados obtidos demonstram que a especialização via *fine-tuning* é um fator relevante para a viabilidade comercial da família Lispector. A família Lispector demonstra maior previsibilidade e alinhamento às correções de referência que modelos generalistas via *prompting*, características desejáveis em sistemas de escrita assistida. O Lispector *small*, com número compacto de parâmetros, alcança desempenho equivalente ao do Lispector *large* e supera, com consistência, modelos maiores sem supervisão — vantagem que se estende também à latência, fator crítico para sistemas de escrita assistida em tempo real. Em termos de latên-

Modelo	Média (ms)	Mediana (ms)	P95 (ms)	Min (ms)	Max (ms)
Fine-tuned (zero-shot)					
ft-lispector-large	3.756	2.040	6.454	1.262	63.676
ft-lispector-small	1.567	1.378	2.601	880	5.435
Prompting (few shot)					
gpt-4.1	2.741	2.556	5.023	1.676	6.789
gpt-4.1-mini	3.277	3.064	5.410	1.853	13.195
gpt-4.1-nano	2.669	2.487	4.033	1.834	4.715
gpt-5	55.015	41.392	162.940	19.265	181.728
gpt-5-mini	21.399	20.792	33.305	8.311	42.597
gpt-5-nano	32.634	31.792	48.946	13.893	76.428

Table 3: Resultados comparativos de latência e tempo de resposta em milissegundos.

cia, em teste com conjunto de dados cujos registros contabilizaram média de 130 tokens, há um contraste severo de latência entre as abordagens, conforme apresentado na Tabela 3. Enquanto o modelo GPT-5 (utilizado via *prompting*) apresenta uma mediana de 41.392 ms e média superior a 55 segundos, tornando-o inviável para interfaces síncronas, o modelo Lispector *small* atinge uma mediana de 1.378 ms. Essa redução drástica de latência permite que a correção ocorra de forma quase instantânea à medida que o usuário digita. O comportamento mais conservador observado nos modelos *fine-tuned* demonstra particular relevância para a experiência do usuário. Os resultados obtidos indicam que o *fine-tuning* favorece correções mais estáveis, pontuais e previsíveis, alinhadas às expectativas de uso em editores de texto. Vale ressaltar que, embora o *fine-tuning* apresente vantagens em desempenho e latência, a abordagem envolve custos iniciais maiores que o *prompting*, incluindo coleta e anotação de dados, treinamento e manutenção do modelo. Para cenários com restrições orçamentárias ou baixo volume de dados disponíveis, o *prompting* pode representar uma alternativa mais acessível. Contudo, os resultados sugerem que, em contextos industriais com demandas de escala e latência, o investimento em *fine-tuning* tende a se justificar. O Lispector está atualmente em operação na plataforma comercial de edição de texto Clarice.ai, processando requisições de usuários reais em escala, o que reforça a viabilidade prática da abordagem proposta.

4.2 Limitações

O conjunto de dados limita-se ao português brasileiro, restringindo a generalização para outras variantes. Além disso, embora os dados englobem uma diversidade natural de gêneros textuais, esta não foi explorada como variável experimental. Outra limitação refere-se à ausência de comparações diretas com ferramentas baseadas em regras, como LanguageTool. A inclusão destes sistemas poderia ampliar a visão de como o Lispector se posiciona no ecossistema de ferramentas de revisão textual. Ademais, os modelos *baseline* foram avaliados com *prompts* padronizados, sem otimização exaustiva; é possível, então, que uma engenharia de *prompt* mais refinada reduzisse parte da diferença observada. A equivalência de desempenho entre Lispector *small* e Lispector *large* indica que modelos compactos são suficientes para a tarefa, embora não permita concluir se *prompts* mais elaborados alcançariam desempenho similar. Ambos os pontos constituem direções relevantes para trabalhos futuros. Por fim, o uso exclusivo de métricas automáticas, embora adequado para comparação em larga escala, não captura com detalhes a aceitabilidade das correções do ponto de vista do usuário final. Reconhecemos que uma análise qualitativa em uma amostra de exemplos poderia enriquecer a interpretação dos resultados, especialmente para identificar casos em que correções válidas mas distintas da referência são penalizadas pelas métricas. Contudo, por se tratar de correção gramatical e ortográfica — e não de revisão estilística —, o espaço de respostas aceitáveis tende a ser mais restrito, o que atenua parcialmente essa limitação.

4.3 Trabalhos futuros

Pretendemos expandir o conjunto de dados com anotações sobre gêneros textuais e contextos comu-

nicativos, como textos acadêmicos, jornalísticos e jurídicos, permitindo investigar estratégias de adaptação mais específicas ao domínio. Outra potencial direção inclui avaliações humanas sistemáticas focadas em critérios de aceitabilidade, fluidez e utilidade percebida pelo usuário. Ademais, estudos em ambientes de produção, incluindo testes A/B com usuários reais, podem fornecer evidências sobre o impacto do *fine-tuning* supervisionado na experiência do usuário.

5 Conclusão

Este trabalho apresentou o Lispector, uma família de modelos de linguagem especializados para revisão gramatical e ortográfica em português brasileiro, desenvolvidos a partir de *fine-tuning* supervisionado em textos reais de usuários. Em uma avaliação comparativa, contrastamos essa abordagem com o uso direto de modelos generalistas de grande escala via *prompting few-shot*, prática amplamente adotada em aplicações industriais. Os resultados demonstram de forma consistente que o *fine-tuning* supervisionado permite ganhos expressivos de desempenho, mesmo em modelos menores. O Lispector *small* superou modelos significativamente maiores utilizados via *prompting*, com ganhos de até 13% em BLEU e 12% em GLEU. Ademais, o Lispector *small* e o Lispector *large* de destacaram no quesito latência. Isso sugere que, para tarefas de correção gramatical e ortográfica, a adaptação via *fine-tuning* oferece vantagens em desempenho e eficiência computacional. O uso atual do Lispector em uma interface de edição de texto reforça a aplicabilidade da abordagem em cenários reais, evidenciando que modelos compactos *fine-tuned* podem atender de forma eficaz às demandas de sistemas industriais de escrita assistida.

6 Agradecimentos

Este trabalho foi financiado pela Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) por meio de bolsas de Treinamento Técnico. Agradecemos à Clarice.ai e a seus usuários, que contribuíram com os dados utilizados.

References

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Christopher Bryant, Zheng Yuan, Muhammad Reza Qorib, Hannan Cao, Hwee Tou Ng, and Ted Briscoe. 2023. Grammatical error correction: A survey of the state of the art. *Computational Linguistics*, page 1–59.

Steven Coyne, Keisuke Sakaguchi, Diana Galvan-Sosa, Michael Zock, and Kentaro Inui. 2023. Analyzing the performance of gpt-3.5 and gpt-4 in grammatical error correction. *Preprint*, arXiv:2303.14342.

Tao Fang, Shu Yang, Kaixin Lan, Derek F. Wong, Jinpeng Hu, Lidia S. Chao, and Yue Zhang. 2023. Is chatgpt a highly fluent grammatical error correction system? a comprehensive evaluation. *Preprint*, arXiv:2304.01746.

Jorge Kinoshita, Laís do Nascimento Salvador, and Carlos Eduardo Dantas de Menezes. 2006. CoGrOO: a Brazilian-Portuguese grammar checker based on the CETENFOLHA corpus. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA).

Ronaldo Teixeira Martins, Ricardo Hasegawa, Maria Das Graças VolpeNunes, Gisele Montilha, and Osvaldo Novais De Oliveira. 1998. Linguistic issues in the development of regra: A grammar checker for brazilian portuguese. *Nat. Lang. Eng.*, 4(4):287–307.

Daniel Naber. 2003. Languagetool.

Courtney Napoles, Keisuke Sakaguchi, Matt Post, and Joel Tetreault. 2015. Ground truth for grammatical error correction metrics. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 588–593, Beijing, China. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Maria Carolina Penteado and Fábio Perez. 2023. Evaluating gpt-3.5 and gpt-4 on grammatical error correction for brazilian portuguese. *Preprint*, arXiv:2306.15788.

Sascha Rothe, Jonathan Mallinson, Eric Malmi, Sebastian Krause, and Aliaksei Severyn. 2022. A simple recipe for multilingual grammatical error correction. *Preprint*, arXiv:2106.03830.

Grounded in Law: A Multi-Stage Anti-Hallucination Pipeline for Legal RAG Systems in Brazilian Portuguese

Arla Figueiredo, João Lucas, Tatiana Ribeiro, Caio Nery, Alan Rios
Caio Hebert, Luiza Florentino, Arthur Silva, Ícaro Feyerabend, Pedro Vidal and Bruno Cabral
Escavador
{arla-figueiredo, joaolucas, tatianaoliveira, caionery, alanrios}
{caiohebert, luizaflorentino, arthursilva, icarofeyerabend}
{pedrovidal, bruno}@escavador.com

Abstract

Large Language Models (LLMs) are effective text generators but create legal citations at non-trivial rates, a failure mode with serious consequences in legal practice. In Brazilian Portuguese the risk is amplified by citation variability (*juridiquês*), fragment-level references (article → paragraph → item), and the need to distinguish jurisdictions and court instances.

We describe EscavAI, a production Retrieval-Augmented Generation (RAG) system deployed at Escavador, a Brazilian legal-technology platform. The system combines (1) domain-tuned hybrid retrieval (lexical, dense, and cross-encoder reranking) over a large-scale legal corpus; (2) grounded generation with explicit citation constraints; and (3) a post-generation *Reference Audit* layer that extracts legislation and jurisprudence mentions via specialized taggers, normalizes them to a canonical schema, checks *existence* against authoritative databases at fragment granularity, verifies *fidelity* against official texts, and triggers targeted rewrites when inconsistencies are detected.

We report production telemetry from 184,895 audited answers containing 43,175 extracted legal references. Legislation references resolve at 81.7%, while jurisprudence references resolve at only 47.1%, identifying case-law normalization as the primary bottleneck for practitioners. Fidelity verification corrected 6.5% of checked answers before delivery, preventing misrepresented legal claims from reaching end users. By converting silent hallucinations into explicit warnings with per-reference status, the system enables legal professionals to trust verified citations and efficiently review flagged ones, rather than manually checking every authority.

1 Introduction

Generative AI can accelerate legal research and drafting, but factual errors in legal text carry severe consequences: citing a non-existent statute, misquoting an article, or inventing a plausible case identifier can lead to professional sanctions. In Brazil, courts have already sanctioned lawyers who submitted AI-generated texts containing fabricated jurisprudence, and recent studies report that specialized legal tools hallucinate in over 17% to 33% of responses (Magesh et al., 2025).

The problem is amplified in Brazilian Portuguese for domain-specific reasons. Legal language (*juridiquês*) contains terms with precise procedural meanings that may confuse multilingual models. Brazil's Civil Law system differs from the common-law concepts that dominate English-heavy pretraining corpora. Brazilian legal citations also require fragment-level precision (articles → paragraphs → subsections → items), making automatic verification harder than validating a single document identifier.

Retrieval-Augmented Generation (RAG) reduces hallucination by grounding output in retrieved documents (Lewis et al., 2020). In legal applications, however, retrieval alone is insufficient: even with the correct law in context, the model can still misquote a fragment, mix up article numbers, or fabricate a case reference. This motivates a *post-generation verification* layer that treats citations as structured objects and validates them against authoritative sources.

We present EscavAI, a production pipeline designed for this setting, and report its behavior over 184,895 audited answers. Our contributions are:

1. A **Reference-Audited RAG** architecture that couples domain-tuned hybrid retrieval with post-generation verification of every extracted legal reference, including existence checks at fragment granularity and fidelity-driven rewriting.
2. A practical **reference extraction and parsing stack** for Brazilian legal text, with taggers for legislation and jurisprudence mentions and a canonical schema for hierarchical fragment paths.
3. A **large-scale production evaluation** over

Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)
Vol. 2, pages 30–34

43,175 references that reveals a pronounced resolution gap between legislation (81.7%) and jurisprudence (47.1%), and shows that fidelity verification catches semantic errors in 6.5% of checked answers.

2 Domain Challenges in Brazilian Law

Two characteristics of the Brazilian legal domain make hallucination mitigation particularly challenging and motivate our design.

Citation complexity. A single provision can appear in many surface forms (e.g., *"Art. 5º, inciso LVII, da CF/88"* or *"artigo quinto, inciso 57 da Constituição Federal"*). References may contain enumerations ("§§ 2º, 3º, 6º e 8º do art. 11"), ranges, and nested paths. A verifier must map these variants to a canonical reference and resolve the cited fragment precisely.

Jurisdictional and instance constraints. For jurisprudence, correct identification depends on court, instance, decision type (*acórdão*, *súmula*, *recurso*), and process identifier formats that vary by tribunal. Models can generate syntactically valid but non-existent case numbers, or cite a valid number under the wrong court.

3 System Architecture

The system follows a four-stage pipeline: **Retrieve** → **Generate** → **Audit** → **Deliver** (Figure 1).

3.1 Stage 1: Hybrid Retrieval

Retrieval is implemented as a multi-phase funnel on a distributed search engine, combining lexical, semantic, and metadata signals. The knowledge base includes jurisprudence and editorial content (news, commentaries) in a vector/lexical index, while legislation and its fragments are served from a dedicated structured store.

Lexical and metadata retrieval. The first phase uses BM25 and field-aware lexical features (nativeRank, fieldMatch) with Portuguese stemming and tuned field weights. Concise legal summaries (*ementa*) receive substantially higher weight than full texts (*inteiro teor*) to limit noise from long documents. Metadata fields (tribunal, state, decision type, date) support fast filtering over millions of documents.

Dense retrieval. We index 768-dimensional embeddings produced by a domain-specific bi-encoder trained on millions of Brazilian legal document pairs using MultipleNegativesRankingLoss with hard negatives mined from top BM25 results. This stage increases recall for paraphrases and synonymy (*"dispensa imotivada"* vs. *"demissão sem justa causa"*), helping retrieval when the same legal idea is expressed with different wording.

Cross-encoder reranking. A quantized ONNX cross-encoder (Nogueira and Cho, 2019) is applied as a global phase on the top candidates. Documents below a confidence threshold of 0.2 are pruned; when no document exceeds this threshold, the system triggers a fallback response.

Query expansion. In the current deployed configuration, each user query is expanded into 7 reformulated sub-queries via function calling. Sub-queries execute in parallel across multiple collections (jurisprudence, *súmulas*, news, commented content), improving coverage of terminological variation.

Domain-specific calibration. Two practical issues required domain-specific adjustments: (i) a non-trivial fraction of jurisprudence entries lack an *ementa* field, requiring dynamic field-weight rebalancing to avoid burying relevant items; and (ii) very short document types (*súmulas*) can dominate lexical scoring due to their density, so we apply type-aware score calibration.

3.2 Stage 2: Grounded Generation

Retrieved documents are formatted into a context window with explicit source identifiers. The generation prompt enforces three constraints: (i) answer using only the retrieved context; (ii) use an explicit Brazilian legal citation style when mentioning authorities; and (iii) refuse when the answer cannot be supported by the available context.

Context budgeting. The system allocates a configurable token budget across entity types. A typical configuration assigns weight 0.50 to case law, 0.40 to commented jurisprudence and legal news, and 0.10 to *súmulas* and commented legislation. Unused budget from one type is redistributed to others, maintaining a total window of 13k to 24k tokens depending on query complexity and active filters.

3.3 Stage 3: Reference Auditing

The core novelty is a post-generation audit pipeline designed to prevent silent citation hallucinations. The audit operates on the *draft answer* and produces: (i) a structured list of audited references with per-reference status; and (ii) a marked-up answer where each reference is linked to the resolved authority when available.

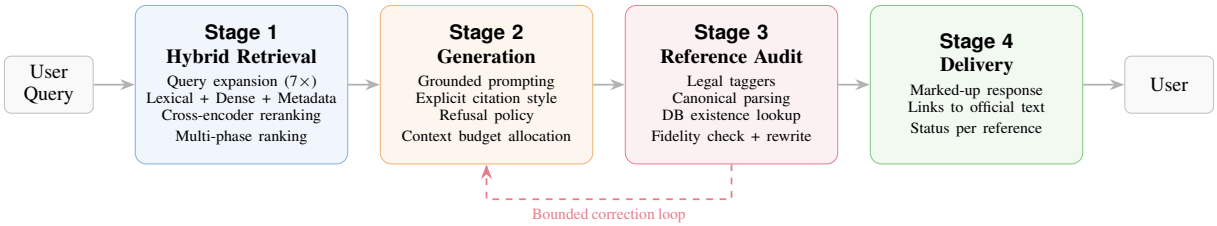


Figure 1: Pipeline overview. The audit stage can trigger a bounded rewrite loop with explicit error context and the retrieved official text for disputed references.

Step 1: Reference extraction. Two specialized taggers identify references in the generated text:

- A **Legislation Tagger** that recognizes mentions of laws, codes, constitutions, decrees, and their fragments. It handles significant surface variation, mapping *"Lei nº 8.666/93"*, *"Lei de Licitações"*, and *"L. 8666"* to the same canonical form.
- A **Jurisprudence Tagger** that identifies court decisions, *súmulas*, and temas, extracting court, decision type, and identifier.

Both taggers model the problem as a sequence labeling task trained on curated data constructed through a produce-consume pipeline: an LLM performs initial annotation, and human annotators validate and correct the output. The taggers are versioned and served via an internal microservice.

Step 2: Canonical parsing. Each tagged span is parsed into a structured object. For legislation, this includes type, number, year, and a fragment path (article → paragraph → inciso → alínea). For jurisprudence, the object includes court, decision type, and case identifier.

Step 3: Existence and fragment resolution. Each canonical object is resolved against authoritative stores: (i) a relational legislation database with fragment-level entries and temporal metadata; and (ii) a search index of case law covering multiple court levels. The audit assigns per-reference status: FOUND, PARTIALLY_FOUND (law exists but a specific fragment is missing), or NOT_FOUND. For legislation, status is also computed per individual fragment.

Step 4: Fidelity verification. When a reference is FOUND, the system compares what the draft answer claims about the authority against the retrieved official text. If the claim is unsupported, the system triggers a targeted rewrite prompt conditioned on the official excerpt, and re-validates the updated answer before delivery.

Table 1: Production telemetry summary at the reference level. RR = fully resolved references.

Metric	Legislation	Jurisprudence	Overall
References	38,914	4,261	43,175
RR%	81.7	47.1	78.2
Unresolved %	12.9	52.9	16.8

3.4 Stage 4: Delivery

When Step 4 triggers a correction, the rewrite loop is bounded (up to 2 iterations) to control latency. The final response is then returned with inline markup tags carrying reference identifiers and status values, plus a structured JSON audit report. The client UI can render links to official texts and highlight unresolved references visually.

4 Evaluation: Production Telemetry

We evaluate the system through automated reference-level audit signals on all audited production answers from launch to February 2026. The dataset contains **184,895 answers** and **43,175 extracted legal references** (38,914 legislation; 4,261 jurisprudence). Our primary goal is to quantify how often generated citations can be resolved to authoritative sources and how often fidelity checks require correction. Table 1 summarizes the reference-level results.

Across all audited answers, 22,595 (12.2%) include at least one explicit legislation or jurisprudence reference extracted by the audit pipeline. These answers account for the 43,175 references summarized in Table 1. Of those answers, 14,860 were fidelity-checked, and 961 were rewritten (6.5% of fidelity-checked answers). The resolution rates in Table 1 are reported at the reference level, whereas the rewrite rate is reported at the answer level.

Two findings are central. First, there is a large gap between legislation and jurisprudence resolution (81.7% vs. 47.1%), indicating that jurisprudence indexing/normalization remains the main bottleneck. Second, fidelity verification is necessary even when references resolve: 6.5% of checked

answers required rewriting, preventing incorrect legal interpretations from reaching end users. Operationally, the audit turns unresolved references into explicit warnings, shifting the failure mode from silent hallucination to transparent uncertainty.

4.1 Qualitative Error Analysis

To complement aggregate telemetry, we manually reviewed representative audit traces and observed four recurrent error classes:

1. **Jurisprudence coverage gaps:** the dominant source of NOT_FOUND case-law references, especially for recently published decisions across many courts.
2. **Long-tail legislation:** unresolved references to municipal, historical, or sparsely indexed norms.
3. **Fragment-level mismatches:** law-level matches where the cited paragraph/item does not resolve, captured as PARTIALLY_FOUND.
4. **Fidelity errors:** incorrect legal claims attached to real authorities, detected and corrected in the rewrite loop.

This analysis reinforces the value of surfacing unresolved references explicitly, rather than silently dropping them. It also points to the main priorities for future iterations: broader jurisprudence coverage and improved fragment-level resolution.

Representative production examples illustrate the behavior:

- *"Súmula 439 do STJ"* → FOUND (correct tribunal and identifier).
- *"§ 4º do art. 3º da Lei 13.105/2015"* → PARTIALLY_FOUND (law resolved, fragment unresolved).
- *"art. 4º da Constituição de 1946"* → NOT_FOUND (outside current coverage).

5 Deployment and Industry Impact

The pipeline is deployed in production at Escavador, a Brazilian legal information platform serving over 14 million monthly visitors. EscavAI supports multiple features: process summarization, process Q&A, general legal Q&A, movement explanation, document drafting, document review, and jurisprudence search, and currently serves over 60,000 monthly active users on AI-powered capabilities. From a compliance perspective, the system operates under Brazil's LGPD with NER-based PII masking. Audit traces provide a governance layer by recording cited authorities, resolution status, and supporting official excerpts.

6 Related Work

RAG is the dominant grounding strategy for LLM systems (Lewis et al., 2020), and strong retrieval stacks often combine hybrid retrieval with reranking (Karpukhin et al., 2020; Nogueira and Cho, 2019). In Brazilian Portuguese, prior legal-NLP resources and language models provide useful foundations for domain adaptation (Luz de Araujo et al., 2018; Souza et al., 2020; Polo et al., 2021; de Mello et al., 2024). For legal assistants, hallucinated authorities remain a key risk (Magesh et al., 2025). Corrective architectures such as Self-RAG and CRAG (Asai et al., 2024; Yan et al., 2024) motivate our domain-specific design choice: explicit post-generation auditing with fragment-level legal resolution and fidelity checking against official sources.

7 Conclusion

We described a production anti-hallucination pipeline for Brazilian Portuguese legal assistants, combining hybrid retrieval with a post-generation Reference Audit layer. Production telemetry over 184,895 answers and 43,175 references reveals that legislation references resolve at 81.7%, while jurisprudence references resolve at only 47.1%, identifying jurisprudence verification as the primary open challenge. Fidelity verification catches semantic errors in 6.5% of checked answers and triggers targeted rewrites before delivery.

The system shifts failures from silent hallucination to explicit uncertainty by flagging unresolved references. Future work includes temporal validity tracking, broader jurisprudence coverage, and improved handling of fragment-level mismatches.

References

A. Asai, Z. Wu, Y. Wang, A. Sil, and H. Hajishirzi. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *Proc. of ICLR*, 2024.

V. Karpukhin, B. Oğuz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, and W. Yih. Dense passage retrieval for open-domain question answering. In *Proc. of EMNLP*, pages 6769–6781, 2020.

P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. Yih, T. Rocktäschel, S. Riedel, and D. Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Proc. of NeurIPS*, volume 33, pages 9459–9474, 2020.

P. H. Luz de Araujo, T. E. de Campos, R. R. R. de Oliveira, M. Stauffer, S. Couto, and P. Bermejo. LeNER-Br: A dataset for named entity recognition in Brazilian legal text. In *Proc. of PROPOR*, pages 313–323, 2018.

V. Magesh, F. Surani, M. Dahl, M. Suzgun, C. D. Manning, and D. E. Ho. Hallucination-free? Assessing the reliability

of leading AI legal research tools. *Journal of Empirical Legal Studies*, 22(1):216–242, 2025.

G. L. de Mello, M. Finger, F. Serras, M. de Mello Carpi, M. M. Jose, P. H. Domingues, and P. Cavalin. PeLLE: Encoder-based language models for Brazilian Portuguese based on open data. In *Proc. of PROPOR*, pages 255–265, 2024.

R. Nogueira and K. Cho. Passage re-ranking with BERT. *arXiv preprint arXiv:1901.04085*, 2019.

F. M. Polo, G. C. F. Mendonça, K. C. J. Parreira, L. Gianvechio, P. Cordeiro, J. B. Ferreira, L. M. P. de Lima, A. C. do Amaral Maia, and R. Vicente. LegalNLP: Natural language processing methods for the Brazilian legal language. In *Proc. of ENIAC*, pages 763–774, 2021.

F. Souza, R. Nogueira, and R. Lotufo. BERTimbau: Pretrained BERT models for Brazilian Portuguese. In *Proc. of BRACIS*, pages 403–417, 2020.

S.-Q. Yan, J.-C. Gu, Y. Zhu, and Z.-H. Ling. Corrective retrieval augmented generation. In *Proc. of ICLR*, 2024.

Socially Responsible and Explainable Automated Fact-Checking and Hate Speech Detection

Francielle Vargas
University of São Paulo
francielleavargas@usp.br

Fabrício Benevenuto
Federal University of Minas Gerais
fabricio@dcc.ufmg.br

Thiago A. S. Pardo
University of São Paulo
taspardo@icmc.usp.br

Abstract

This Ph.D. dissertation advances the state-of-the-art in Natural Language Processing (NLP) for Portuguese by proposing new and innovative data resources and explainable methods for hate speech detection and automated fact-checking. The thesis introduces several benchmark datasets for Brazilian Portuguese, HateBR, HateBRXplain, HateBRMoralXplain, MFTCXplain, MOL, and FactNews, which have been widely adopted by the research community and address critical gaps in the availability of high-quality annotated resources for Portuguese. In addition, this dissertation proposes novel post-hoc and self-explaining NLP methods: Sentence-Level Factual Reasoning (SELFAR), Social Stereotype Analysis (SSA), Contextual Bag-of-Words with Interpretable Input and Feature Optimization (B+M), Supervised Rational Attention (SRA), and Supervised Moral Rational Attention (SMRA). Across multiple tasks and datasets in Portuguese, these methods outperform baselines while improving interpretability and robustness, demonstrating that explainability and performance can be jointly optimized. Finally, this thesis has achieved significant national and international impact, being cited by leading universities and research institutes worldwide and fostering new M.Sc. and Ph.D. research projects in Brazil. Its scientific and social contributions have also been recognized with multiple prestigious national and international awards, including the Google LARA, the Maria Carolina Monard Best Thesis Award in Artificial Intelligence, the Trevisan Prize for Students "AI for Good" from Bocconi University for rigorous computer science research in AI with social impact, and the Diversity and Inclusion Award from the Association for Computational Linguistics (ACL). Lastly, this thesis has received two nominations for the Brazilian Computer Society Thesis Awards in Computer Science, and in Multimedia, Hypermedia, and Web.

1 Introduction

Although the proliferation of misinformation and hate speech is a global challenge, most existing fact-checking and hate speech detection models remain focused on English and rely on opaque architectures. As a result, these "black-box" approaches fail to provide meaningful rationales for their predictions, while Portuguese continues to lack high-quality corpora, benchmarks, and explainable methods. This lack of transparency introduces significant risks, including biased and unreliable model behavior, which has become a major concern in the Artificial Intelligence (AI) field (May et al., 2019). These limitations reveal a clear research gap, constraining both scientific progress and the development of transparent, effective, and culturally aware solutions for Portuguese language processing.

Hate speech detection technologies often inherit biases from their training data (Davidson et al., 2019), which may reflect subjective human annotations (Al Kuwatly et al., 2020; Sap et al., 2022; Vargas et al., 2022). These biases can reinforce social discrimination, such as racial and gender biases, especially when deployed at scale (Davani et al., 2023; Vargas et al., 2023a; Chuang et al., 2021; Sap et al., 2019; Davidson et al., 2019). Table 1 evidences this issue, showing two documents classified as hate speech by a fine-tuned BERT classifier (Kennedy et al., 2020). Note that while the second document from Gab social media contains explicit hate speech, the first document, extracted from the New York Times[1], does not. However, the fine-tuned BERT classifier incorrectly classified both as hate speech, due to the presence of group identifiers such as "black' and "Africans."

In similar settings, the issue of bias in fact-checking has been extensively analyzed in the literature (Park et al., 2021; Baly et al., 2018; Soprano et al., 2024; Vargas et al., 2023c). Biases

[1] https://www.nytimes.com/

Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)
Vol. 2, pages 35–42

Documents	Predicted Class
For many Africans, the most threatening kind of ethnic hatred is black against black. - New York Times	hate speech
There is a great discrepancy between whites and blacks in SA. It is … [because] blacks will always be the most backward race in the world - Anonymous user, Gab.com	hate speech

Table 1: Two documents classified as hate speech by a fine-tuned BERT classifier. Group identifiers are underlined (Kennedy et al., 2020).

N.	Claims	Rate
1	"Latina workers make 54 cents for every dollar earned by white, non-Hispanic men" - Democratic Senator (tweet)	True
2	"A proposal in Syracuse would pay gang members $100-$200 per week to stay out of trouble" - Republican state legislator from New York (tweet)	Mostly True

Table 2: Examples of manually fact-checked claims and their assigned ratings published by PolitiFact.

(e.g., media bias, political bias,) in automated fact-checking may be also introduced during data training. Fact-checking models tend to rely on these biases without fully learning the underlying task. Instead, they often learn misleading correlations between news patterns and veracity labels as simplifications, rather than integrating the information to reason effectively (Wu et al., 2022). As a result, these models may not only fail when applied to real-life situations, where news patterns vary widely, but they can also undermine public trust and exacerbate political polarization (Kuzmin et al., 2020). For example, prior studies assessing the performance of human fact-checkers have reported conflicting findings (Amazeen, 2015) and identified significant inconsistencies among major fact-checking organizations such as PolitiFact, The Fact Checker[2], and FactCheck.org in their evaluations of statements on topics such as climate change, racism, and national debt (Marietta et al., 2015). In addition, only 10% of statements were fact-checked by both organizations in their study, with agreement primarily observed for statements classified as clearly true or false, but significantly lower agreement for ambiguous claims that highlight the inherent subjectivity in the manual assignment of veracity ratings, raising concerns about potential biases such as selective claim verification and inconsistencies in evaluation criteria (Nieminen and Rapeli, 2019). An example of this lack of precision in defining veracity ratings is shown in Table 2. Note that in the second example, the label "mostly true" is assigned despite the low consistence of the evidence.

Moreover, media and political biases can be introduced during data collection and training, influencing the behavior of fact-checking models. Consequently, rather than learning the fundamental task of factual verification, these models may inadvertently internalize misleading correlations between news patterns and veracity labels as shortcuts (Wu et al., 2022). This reliance on biased patterns not only compromises model performance in real-world scenarios, where news structures vary significantly, but also poses broader societal risks, including diminished public trust in fact-checking and increased political polarization (Kuzmin et al., 2020). Thus, ensuring the transparency and accountability of automated fact-checking systems is both a technical necessity and a social imperative. To address these challenges, fact-checking models should either incorporate post-hoc explanations for their outputs or embed interpretability mechanisms directly within their architecture (Kotonya and Toni, 2020).

Therefore, the inability of NLP models to provide rationales for their decisions remains a significant barrier to their broader adoption (Gongane et al., 2024). In the context of automated fact-checking and hate speech detection, this lack of transparency raises serious ethical concerns regarding model reliability and fairness. In response to these critical issues, this thesis advances transparent and explainable approaches for Portuguese language processing, with a particular focus on fact-checking and hate speech detection, two tasks that are central to combating misinformation and sustaining a fair and democratic society. Specifically, this Ph.D. thesis introduced several benchmark datasets for Portuguese (e.g., HateBR, HateBRXplain, HateBRMoralXplain, MFTCXplain, FactNews, and MOL), and developed new post-hoc and self-explaining computational methods (e.g., SELFAR, SSA, B+M, SRA, SMRA) to ensure that both data and models are explainable and socially

[2] https://www.washingtonpost.com/politics/fact-checker/

aligned. Notably, over multiple tasks in Portuguese, these methods reliably outperform the baselines and simultaneously improve interpretability and robustness, significantly contributing to advance the state-of-the-art in the computational processing of Portuguese.

2 Theoretical Background

Explainable Artificial Intelligence (XAI) systems can explain their reasoning to human users and express knowledge about how they will behave in the future (Guidotti et al., 2018; Adadi and Berrada, 2018). In this context, XAI methods provide the causes of a single prediction, a set of predictions, or all predictions of a model by identifying the input, model, or training data parts that most influence the model's outcome (Balkir et al., 2022). The key concept in explainability involves the types of explanations, often categorized in literature into two main groups: (i) local and global explanations and (ii) self and post-hoc explanations (Guidotti et al., 2018; Adadi and Berrada, 2018), as follow.

Local explanations: This first type of explanation provides information or justification for the model's prediction regarding a specific input. Furthermore, local explanations are also known as model-agnostic explanations, meaning they do not consider the structure of the model.

Global explanations: This second type of explanation arises directly from the prediction process. It provides justification by revealing how the model's predictive mechanism works, regardless of any specific input. Moreover, global explanations are also known as model-specific explanations, as they consider the internal structure of the model's process and rely on the specific architecture used.

In addition, explanations differ based on whether they are generated as part of the prediction process or require post-processing after the model makes a prediction. These are categorized as self-explaining and post-hoc explaining, as described below:

Self-explaining: This type of approach is also referred to as directly interpretable (Arya et al., 2019). It generates explanations simultaneously with predictions, utilizing information provided by the model during the prediction process. For example, decision trees and rule-based models exemplify global self-explaining models, whereas feature saliency approaches, such as attention mechanisms or feature engineering, serve as examples of local self-explaining models.

Post-hoc explaining: This approach generates explanations after the model has been built, requiring an additional operation performed after predictions are made. LIME (Local Interpretable Model-Agnostic Explanations) (Ribeiro et al., 2016) and SHAP (SHapley Additive exPlanations) (Lundberg and Lee, 2017) are examples of post-hoc explaining methods. LIME provides local explanations for predictions by perturbing input data and observing the resulting changes in the model's predictions. In contrast, SHAP measures the contribution of each feature to the prediction by considering all possible combinations of features and can be used for both local and global explanations.

3 Problem, Motivation and Objectives

Advances in NLP models have led to a decline in transparency, increasing risks such as bias and reduced accountability, particularly in socially sensitive tasks such as fact-checking and hate speech detection. These challenges are particularly notable for Portuguese, which still lacks adequate resources and explainable methods. Hence, the main objectives of this thesis are: (i) to analyze the risks, and limitations of "black-box" NLP models applied to fact-checking and hate speech detection in Portuguese; (ii) to design and release benchmark datasets for Portuguese, enriched with expert annotations and human-interpretable rationales; (iii) to propose novel explainable NLP methods that integrate interpretability into the learning process while maintaining state-of-the-art predictive performance; and (iv) to contribute to the advancement of the Portuguese NLP community by providing open-access data, methods, and evaluation metrics.

4 Research Hypotheses

This thesis was guided by two main hypotheses. First, research on explainability is essential for building trust, ensuring fairness, and promoting the responsible use of AI, particularly for understanding the risks and social impacts of "black-box" NLP models for automated fact-checking and hate speech detection in Portuguese. Second, by revealing when models rely on spurious patterns or social biases, explainability supports error diagnosis, improves robustness, and enables accountability, which is crucial in hate speech detection and fact-checking, where understanding the reasons behind a decision is as important as the decision itself.

5 Research Methodology

The research adopts a data-driven and evaluation-oriented methodology, comprising: (i) the construction of benchmark datasets for Portuguese with expert annotations and rationale-level supervision; (ii) the development of explainable NLP models, including self-explaining and post-hoc architectures; (iii) the evaluation of models in Portuguese language using standard NLP metrics alongside explainability and fairness-oriented assessments; and (iv) a comparative analysis against strong state-of-the-art baselines to quantify advances in performance, transparency, and robustness.

6 Research and Social Impact on Portuguese Language Processing

The research presented in this dissertation has had a significant national and international impact on the computational processing of Portuguese, particularly in the context of socially-aware NLP tasks such as hate speech detection, misinformation, and explainable AI. The proposed methodologies and language resources directly address theoretical and technological challenges, with a special focus on Brazilian Portuguese, contributing to the advancement of robust, transparent, and socially responsible Portuguese-language NLP systems. A major outcome of this work is the creation of widely adopted benchmark datasets for Portuguese, including the **HateBR**,**HateBRXplain**, **HateBRMoralXplain**, **FactNews**, **Multilingual Offensive Lexicon (MOL)**, and **MFTCXplain** all of which have been extensively used for benchmarking, evaluation, and methodological development in NLP research. These resources fill critical gaps in high-quality, expert-annotated datasets for Portuguese. In the **international context**, the impact of this dissertation is evidenced by its adoption and citation by leading research institutions and universities such as Microsoft Research, Carnegie Mellon University, Harvard University, University of Maryland, University of Turin, Technical University of Munich, University of Bonn, University of Lisbon, National University of Singapore, Vrije Universiteit Amsterdam, and the Rochester Institute of Technology. These institutions have leveraged the proposed datasets, as well as our methods, such **B+M, SSA, SELFAR, SRA, SMRA**, for advancing multilingual and Portuguese-centered NLP research. Her international recognition continues to grow. In December 2025, **Dr. Vargas was invited to present her research at the highly prestigious ASML Synthesizer Open Showcase 2025**[3] **at Harvard University**, where she presented "*Brazil #WithoutHate: Self-Explaining and Moral-Aware AI for Hate Speech Detection*." This invited presentation highlights the maturity, originality, and societal relevance of her doctoral research, as well as its contribution to the development of explainable and fair AI systems for the processing of Portuguese. In the **national context**, this work has directly influenced graduate-level research in Brazil. Several universities, including UFMG, USP, UFF, UFCG, UDESC, UFOP, etc., have proposed or developed MSc and PhD theses centered on hate speech and misinformation analysis using the datasets and methodologies introduced in this dissertation. The relevance and visibility of this research also extend beyond publications. The author was invited to serve as a visiting researcher at the University of Southern California, and to present this work at Leibniz Institute for the Social Sciences. Additionally, she has actively contributed to international NLP research communities through service roles, including participation in the organizing committees of ICWSM (2021, 2022, 2023) and WOAH (2025) and DeepXplain (2025), as well as program committee in ACL, EMNLP, NAACL, LREC, and COLING. Overall, this dissertation contributes novel methods, datasets, and evaluation frameworks for Portuguese, promote reproducible research, and support the development of linguistically and socially responsible NLP technologies, fully aligning with the objectives of the PROPOR Best PhD Dissertation Award.

7 Thesis Outcomes

The thesis generated several outcomes, including:

1. **15 (fifteen) published papers in top-tier international conferences**: In total, 15 (fifteen) papers were published in top-tier AI and NLP international conferences, workshops, and journals (e.g., EMNLP, NAACL, LREC, RANLP, NLP journal, AAAI, etc.) (see list of publised papers).

2. **Several benchmark datasets for Portuguese**: HateBR[4] (Vargas et al., 2022), HateBRXplain (Salles et al., 2025), and HateBRMoralXplain (Vargas et al., 2026) (the first

[3] https://cyber.harvard.edu/events/asml-2025-synthesizer

[4] https://github.com/franciellevargas/HateBR

large-scale expert-annotated corpora for hate speech detection in Brazilian Portuguese, including hate speech and moral rationales for explainability); MOL[5] (Vargas et al., 2024a) (the first multilingual offensive lexicon, extracted from the HateBR containing 1,000 explicit and "clues" to identify implicit terms in Portuguese, translated into five languages while accounting for cultural aspects.); FactNews[6] (Vargas et al., 2023c) (a sentence-level annotated dataset for fact-checking in Portuguese); and MFTCXplain[7] (Trager et al., 2025), the first benchmark for evaluating moral reasoning of LLMs.

3. **5 (Five) new post-hoc and self-explaining computational methods evaluated for Portuguese**: SELFAR[8] (Vargas et al., 2024b) (the first explainable fact-checking method in Portuguese); SSA[9] (Vargas et al., 2023a) (a counter-stereotype post-hoc explanation method to assess social bias in hate speech classifiers); B+M[10] (Vargas et al., 2021) (a contextual BoW with interpretable input and feature optimization for explainable hate speech detection); and SRA (Eilertsen et al., 2025) and SMRA[11] (Vargas et al., 2026) (self-explaining methods that integrating hate speech and moral human-annotated rationales into deep learning models by attention alignment loss).

4. **A Web platform to combat hate speech in Portuguese**: NoHateBrazil[12] (Vargas et al., 2023b), a public platform for hate speech detection in Brazilian Portuguese.

5. **Microsoft leveraged the HateBR dataset to train LLMs**: Microsoft Research has used the proposed HateBR dataset to train two 2 (two) LLMs: CultureLLM (Li et al., 2024) and CulturePark (Li et al., 2025), demonstrating its clear relevance and practical applicability.

6. **National and international awards**: This work has been widely recognized through prestigious national and international awards and grants, including the Google LARA[13], the Maria Carolina Monard Award Best Thesis in AI[14], International Trevisan Prize for Students "AI for Good"[15], and was nominated for the Brazilian Computer Society's Thesis Awards in Computer Science (Vargas et al., 2025a) and Multimedia, Hypermedia and Web (Vargas et al., 2025b). Finally, the author of this thesis received a Diversity & Inclusion (D&I) award from the Association for Computational Linguistics (ACL) for EMNLP and NAACL 2024. This award is granted to Ph.D. students in recognition of their outstanding contributions and achievements in Natural Language Processing and Computational Linguistics, as well as to applicants from underrepresented groups presenting a paper at the main conference.

7. **Invited as an international visiting researcher**: The author of this thesis was invited to visit the University of Southern California (USC) and to present her work at an event of the Applied Social Media Lab at the *Berkman Klein Center for Internet & Society*, **Harvard University**.

8. **Invited as an international keynote speaker**: The author of this thesis served as a keynote speaker at the Conference on Harmful Online Communication at the Leibniz Institute for the Social Sciences (GESIS), in Germany, and at the Conference cum Conclave on Emerging Trends in Journalistic and Media Practices at DG Vaishnav College, in India.

9. **Invited roundtable lead at ICLR 2026 workshop**: The author of this thesis was invited to serve as a Roundtable Lead at the *Algorithmic Fairness Across Alignment Procedures and Agentic Systems (AFAA) Workshop*, held at the International Conference on Learning Representations 2026[16].

10. **Invited to top-tier NLP international conference and journals program committees**:

[5] https://github.com/franciellevargas/MOL
[6] https://github.com/franciellevargas/FactNews
[7] https://github.com/franciellevargas/MFTCXplain
[8] https://github.com/franciellevargas/SELFAR
[9] https://github.com/franciellevargas/SSA
[10] https://aclanthology.org/2021.ranlp-1.161/
[11] https://github.com/franciellevargas/SMRA
[12] http://143.107.183.175:14581/
[13] https://research.google/programs-and-events/phd-fellowship/
[14] https://www.icmc.usp.br/institucional/premios/premio-maria-carolina-monard
[15] https://cs.unibocconi.eu/news/trevisan-prize-students-ai-good-winners
[16] https://www.afciworkshop.org/afaa-2026

The author of this thesis was invited to serve on the program committees of Language Resources and Evaluation, Expert Systems with Applications, Online Social Networks and Media, Natural Language Processing Journals, as well as EMNLP, ACL, NAACL, RANLP, COLING, LREC, ICWSM, WWW and CIKM conferences.

11. **Co-organizer of top-tier international conferences and workshops**: The author of this thesis co-organized the prestigious ICWSM conference in 2021, 2022 and 2023, and was selected to organize ACL Workshop on Online Abuse and Harms (WOAH 2025) (Calabrese et al., 2025), and proposed the IJCNN 2025 Special Session on Explainable Deep Neural Networks for Responsible AI: Post-Hoc and Self-Explaining Approaches (DeepXplain 2025) [17].

12. **+10 new Ph.D. and M.Sc. theses, and undergraduate research projects in Brazil**: Several prestigious public universities in Brazil have used our resources to propose new Ph.D. and M.Sc. theses, including institutions such as USP, UFMG, UFF, UFOP, UFC, among others.

13. **Co-advisor of a computer science master's student with a published paper in a top-tier NLP venue**: The author of this thesis co-advised a master's student in computer science at DCC-UFMG, whose work was published at COLING 2025, a top-tier international NLP conference (see paper).

14. **High citations from prestigious international institutions**: Research by prestigious institutions (e.g. Carnegie Mellon University, University of Maryland, Harvard University), has been significantly influenced by the resources provided in this thesis, as reflected in a substantial number of citations (Google Scholar: 325 citations) [18].

15. **Collaboration with researchers across 10 countries on 5 continents**: The author of this thesis has led projects and collaborated with researchers in ten different countries on five continents (see research projects).

16. **Press Coverage**: The results of this dissertation received relevant media coverage, with nine reports, including one international outlet (see).

[17] https://deepxplain.github.io/

[18] https://tinyurl.com/2ckwn6vd

8 Ph.D. Dissertation

Link: https://tinyurl.com/7ps8ykys

9 Conclusion

This Ph.D. dissertation advances the state-of-the-art in Natural Language Processing for Portuguese by addressing the lack of high-quality resources and explainable methods for socially sensitive tasks such as hate speech detection and automated fact-checking. To bridge this gap, this thesis introduced several benchmark datasets for Brazilian Portuguese with expert annotations and rationale-level supervision. In addition, it proposed novel explainable computational methods that integrate post-hoc and self-explaining approaches to improve transparency in classical and neural machine learning models. Experimental results across multiple tasks demonstrate that explainability and predictive performance can be jointly optimized, leading to more robust and trustworthy AI systems. Beyond its methodological contributions, this research has had a significant scientific and societal impact. The datasets and methods proposed in this thesis have been widely adopted by the international research community, enabling new studies and graduate research projects while strengthening Portuguese-centered NLP. Overall, this dissertation contributes new resources and methods that advance explainable and socially responsible NLP for Portuguese, reinforcing the importance of transparency and accountability in modern AI and NLP systems.

Acknowledgements

The authors are grateful to São Paulo Research Foundation – FAPESP (grant #2025/01118-2) for financial support. This work had the support of the Ministry of Science, Technology and Innovation, with resources of Law N. 8,248, of October 23, 1991, within the scope of PPI-SOFTEX, coordinated by Softex and published as Residence in TIC 13, DOU 01245.010222/2022-44. It was also supported by Instituto Nacional de Ciência e Tecnologia em Inteligência Artificial Responsável para Linguística Computacional, Tratamento e Disseminação de Informação (INCT-TILD-IAR).

References

Amina Adadi and Mohammed Berrada. 2018. A survey on explainable artificial intelligence (xai). *IEEE Access*, 6:52138–52160.

Hala Al Kuwatly, Maximilian Wich, and Georg Groh. 2020. Identifying and measuring annotator bias based on annotators' demographic characteristics. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 184–190, Held Online.

Michelle Amazeen. 2015. Revisiting the epistemology of fact-checking. *Critical Review*, 27(1):1–30.

Vijay Arya, Rachel K. E. Bellamy, Pin-Yu Chen, Amit Dhurandhar, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Q. Vera Liao, Ronny Luss, Aleksandra Mojsilovic, Sami Mourad, Pablo Pedemonte, Ramya Raghavendra, John T. Richards, Prasanna Sattigeri, Karthikeyan Shanmugam, Moninder Singh, Kush R. Varshney, Dennis Wei, and Yunfeng Zhang. 2019. One explanation does not fit all: A toolkit and taxonomy of ai explainability techniques. *CoRR*, abs/1909.03012.

Esma Balkir, Svetlana Kiritchenko, Isar Nejadgholi, and Kathleen Fraser. 2022. Challenges in applying explainability methods to improve the fairness of NLP models. In *Proceedings of the 2nd Workshop on Trustworthy Natural Language Processing*, pages 80–92, Seattle, USA.

Ramy Baly, Georgi Karadzhov, Dimitar Alexandrov, James Glass, and Preslav Nakov. 2018. Predicting factuality of reporting and bias of news media sources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3528–3539, Brussels, Belgium.

Agostina Calabrese, Christine de Kock, Debora Nozza, Flor Miriam Plaza-del Arco, Zeerak Talat, and Francielle Vargas, editors. 2025. *Proceedings of the The 9th Workshop on Online Abuse and Harms (WOAH)*. Association for Computational Linguistics, Vienna, Austria.

Yung-Sung Chuang, Mingye Gao, Hongyin Luo, James Glass, Hung-yi Lee, Yun-Nung Chen, and Shang-Wen Li. 2021. Mitigating biases in toxic language detection through invariant rationalization. In *Proceedings of the 5th Workshop on Online Abuse and Harms*, pages 114–120, Held Online.

Aida Mostafazadeh Davani, Mohammad Atari, Brendan Kennedy, and Morteza Dehghani. 2023. Hate speech classifiers learn normative social stereotypes. *Transactions of the Association for Computational Linguistics*, 11:300–319.

Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. 2019. Racial bias in hate speech and abusive language detection datasets. In *Proceedings of the 3rd Workshop on Abusive Language Online*, pages 25–35, Florence, Italy.

Brage Eilertsen, Røskva Bjørgfinsdóttir, Francielle Vargas, and Ali Ramezani-Kebrya. 2025. Aligning attention with human rationales for self-explaining hate speech detection. *Preprint*, arXiv:2511.07065.

Vaishali U. Gongane, Mousami V. Munot, and Alwin D. Anuse. 2024. A survey of explainable AI techniques for detection of fake news and hate speech on social media platforms. *Journal of Computational Social Science*, 7(1):587–623.

Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2018. A survey of methods for explaining black box models. *ACM Comput. Surv.*, 51(5).

Brendan Kennedy, Xisen Jin, Aida Mostafazadeh Davani, Morteza Dehghani, and Xiang Ren. 2020. Contextualizing hate speech classifiers with post-hoc explanation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5435–5442, Held Online.

Neema Kotonya and Francesca Toni. 2020. Explainable automated fact-checking: A survey. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5430–5443, Barcelona, Spain.

Gleb Kuzmin, Daniil Larionov, Dina Pisarevskaya, and Ivan Smirnov. 2020. Fake news detection for the Russian language. In *Proceedings of the 3rd International Workshop on Rumours and Deception in Social Media*, pages 45–57, Barcelona, Spain.

Cheng Li, Mengzhuo Chen, Jindong Wang, Sunayana Sitaram, and Xing Xie. 2024. Culturellm: Incorporating cultural differences into large language models. In *Advances in Neural Information Processing Systems*, volume 37, pages 84799–84838. Curran Associates, Inc.

Cheng Li, Damien Teney, Linyi Yang, Qingsong Wen, Xing Xie, and Jindong Wang. 2025. Culturepark: boosting cross-cultural understanding in large language models. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, Red Hook, NY, USA.

Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.

Morgan Marietta, David C. Barker, and Todd Bowser. 2015. Fact-checking polarized politics: Does the fact-check industry provide consistent guidance on disputed realities? *The Forum*, 13(4):577–596.

Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. 2019. On measuring social biases in sentence encoders. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 622–628, Minneapolis, Minnesota.

Sakari Nieminen and Lauri Rapeli. 2019. Fighting misperceptions and doubting journalists' objectivity: A review of fact-checking literature. *Political Studies Review*, 17(3):296–309.

Sungkyu Park, Jaimie Yejean Park, Jeong-han Kang, and Meeyoung Cha. 2021. The presence of unexpected biases in online fact-checking. *Harvard Kennedy School Misinformation Review*, 2(1).

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144.

Isadora Salles, Francielle Vargas, and Fabrício Benevenuto. 2025. HateBRXplain: A benchmark dataset with human-annotated rationales for explainable hate speech detection in Brazilian Portuguese. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6659–6669, Abu Dhabi, UAE.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678, Florence, Italy.

Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A. Smith. 2022. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 5884–5906, Seattle, United States.

Michael Soprano, Kevin Roitero, David La Barbera, Davide Ceolin, Damiano Spina, Gianluca Demartini, and Stefano Mizzaro. 2024. Cognitive biases in fact-checking and their countermeasures: A review. *Inf. Process. Manage.*, 61(3).

Jackson Trager, Francielle Vargas, Diego Alves, Matteo Guida, Mikel K. Ngueajio, Ameeta Agrawal, Yalda Daryani, Farzan Karimi Malekabadi, and Flor Miriam Plaza-del Arco. 2025. MFTCXplain: A multilingual benchmark dataset for evaluating the moral reasoning of LLMs through multi-hop hate speech explanation. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 15709–15740, Suzhou, China.

Francielle Vargas, Isabelle Carvalho, Ali Hürriyetoğlu, Thiago Pardo, and Fabrício Benevenuto. 2023a. Socially responsible hate speech detection: Can classifiers reflect social stereotypes? In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 1187–1196, Varna, Bulgaria.

Francielle Vargas, Isabelle Carvalho, Thiago A. S. Pardo, and Fabrício Benevenuto. 2024a. Context-aware and expert data resources for brazilian portuguese hate speech detection. *Natural Language Processing*, 31(2):435–456.

Francielle Vargas, Isabelle Carvalho, Fabiana Rodrigues de Góes, Thiago Pardo, and Fabrício Benevenuto. 2022. HateBR: A large expert annotated corpus of Brazilian Instagram comments for offensive language and hate speech detection. In *Proceedings of the 13th Language Resources and Evaluation Conference*, pages 7174–7183, Marseille, France.

Francielle Vargas, Isabelle Carvalho, Wolfgang Schmeisser-Nieto, Fabrício Benevenuto, and Thiago Pardo. 2023b. NoHateBrazil: A Brazilian Portuguese text offensiveness analysis system. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 1180–1186, Varna, Bulgaria.

Francielle Vargas, Kokil Jaidka, Thiago Pardo, and Fabrício Benevenuto. 2023c. Predicting sentence-level factuality of news and bias of media outlets. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 1197–1206, Varna, Bulgaria.

Francielle Vargas, Thiago Pardo, and Fabrício Benevenuto. 2025a. Socially responsible and explainable automated fact-checking and hate speech detection. In *Anais do XXXVIII Concurso de Teses e Dissertações*, pages 75–84, Porto Alegre, RS, Brasil. SBC.

Francielle Vargas, Thiago Pardo, and Fabrício Benevenuto. 2025b. Socially responsible and explainable automated fact-checking and hate speech detection. In *Anais Estendidos do XXXI Simpósio Brasileiro de Sistemas Multimídia e Web*, pages 25–26, Porto Alegre, RS, Brasil. SBC.

Francielle Vargas, Fabiana Rodrigues de Góes, Isabelle Carvalho, Fabrício Benevenuto, and Thiago Pardo. 2021. Contextual-lexicon approach for abusive language detection. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing*, pages 1438–1447, Held Online.

Francielle Vargas, Isadora Salles, Diego Alves, Ameeta Agrawal, Thiago A. S. Pardo, and Fabrício Benevenuto. 2024b. Improving explainable fact-checking via sentence-level factual reasoning. In *Proceedings of the Seventh Fact Extraction and VERification Workshop*, pages 192–204, Miami, USA.

Francielle Vargas, Jackson Trager, Diego Alves, Surendrabikram Thapa, Matteo Guida, Berk Atil, Daryna Dementieva, Andrew Smart, and Ameeta Agrawal. 2026. Self-explaining hate speech detection with moral rationales. *Preprint*, arXiv:2601.03481.

Junfei Wu, Qiang Liu, Weizhi Xu, and Shu Wu. 2022. Bias mitigation for evidence-aware fake news detection by causal intervention. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, page 2308–2313, New York, USA.

Automated Essay Scoring for Brazilian Portuguese
Evidence from Cross-Prompt Evaluation of ENEM Essays

André Barbosa and Denis Deratani Mauá
Institute of Mathematics and Statistics, University of São Paulo, São Paulo, Brazil
{andre.barbosa, ddm}@ime.usp.br

Abstract

Brazil's ENEM, a high-stakes assessment determining university admission for millions of students annually, creates an immense evaluation burden where human raters process hundreds of essays daily. Automated Essay Scoring (AES) offers a potential solution, yet Portuguese-language systems remain understudied due to fragmented datasets and the complexity of ENEM's multi-trait rubric. This work investigated cross-prompt, trait-specific essay scoring using a corpus of 385 essays across 38 prompts, where models evaluated essays on unseen prompts across five traits scored on a six-point ordinal scale. We compared three model classes: feature-based methods (72 features), encoder-only transformers (109M–1.5B parameters), and decoder architectures (2.4B–671B parameters) with fine-tuned and zero-shot configurations. Experiments under varying information access and rubric conditioning revealed that no single approach serves all evaluation needs: encoder models excel at mechanical traits (fluency, cohesion) despite context limitations; decoder models achieve superior performance on argumentation (QWK 0.73) and writing style (QWK 0.60) when provided full context; and language-specific pretraining benefits only surface-level features without improving complex reasoning. Best-performing models achieved QWK scores of 0.60–0.73. Gaps to oracle bounds ranged from 0.15 (argumentation) to 0.29 (writing style), with the largest disparities in writing style and persuasiveness.

1 Introduction

Automatic Essay Scoring (AES) systems promise to release educators from the burden of grading written assignments, scaling up the ability to provide timely, consistent, and useful feedback (Page, 1966). These systems have matured significantly, evolving from feature engineering approaches (Page, 1966; Attali and Burstein, 2006; Attali, 2013) to deep neural networks (Taghipour and Ng, 2016; Dong et al., 2017; Alikaniotis et al., 2016) and, more recently, to architectures that leverage Large Language Models (Rodriguez et al., 2019; Mansour et al., 2024). However, the vast majority of this progress focuses on English corpora, leaving Portuguese-language systems understudied.

Brazil's ENEM (*Exame Nacional do Ensino Médio*) exemplifies this challenge. The high-stakes examination annually evaluates 3.9 million students, serving as the primary gateway to higher education. Human raters process 100–200 essays daily under a 20-day evaluation window, with students waiting up to 8 weeks for official results.[1] Secondary schools face compounding difficulties: class sizes of 30–50 students, limited faculty allocation for essay review, and insufficient resources for ENEM-style assessment. The consequences are predictable: delayed feedback impedes learning, and teacher burnout affects evaluation quality.

Developing AES systems to address these challenges faces its own obstacles. Existing datasets suffer from parsing artifacts and lack data provenance (Marinho et al., 2021). Prior empirical work has been limited to feature-based methods or shallow neural networks (Amorim and Veloso, 2017; Fonseca et al., 2018), with no systematic analysis of modern transformer-based architectures. Additionally, most research focuses on single-prompt scoring, which does not reflect realistic deployment where systems must generalize to unseen topics.

This work addressed four research questions. Firstly, what information do different traits require, that is, do all traits benefit equally from access to essay prompts and supporting texts? Secondly, how does information access (prompt-blind vs. prompt-aware) affect performance across the five ENEM traits. Thirdly, do Portuguese-specific models outperform multilingual alternatives, and for which

[1] As reported in https://tinyurl.com/57e5xsdv

Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)
Vol. 2, pages 43–48
April 13-16, 2026

traits? Lastly, what are the practical trade-offs between model architectures regarding accuracy, computational cost, and inference latency?

We hypothesized that different traits require different computational approaches: mechanical traits such as fluency and cohesion can be evaluated with limited context, while argumentative quality requires access to the essay prompt and supporting materials. To address these questions, this work provided an extensive empirical analysis comparing 15 models spanning three paradigms: feature-based methods (72 linguistic features), encoder-only transformers (109M–1.5B parameters), and decoder architectures (2.4B–671B parameters) including fine-tuned and zero-shot configurations.

The main contributions of this dissertation can be outlined as follows:

1. A validated benchmark corpus of 385 essays across 38 prompts with expert annotations from two independent graders;
2. A systematic comparison of model architectures under varying information access and rubric conditioning;
3. A formal framework distinguishing prompt-blind and prompt-aware scoring with three rubric strategies (Student, Mixed, Grader);
4. Trait-specific analysis demonstrating that encoder models excel at mechanical traits despite context limitations while decoder models achieve superior performance on argumentation (QWK 0.73) and style (QWK 0.60) when provided full context.
5. Open research artifacts including the annotated corpus and evaluation scripts. The dataset and models generated during this research are available in `https://tinyurl.com/245mxct9`. Experiments and code used are available in `https://github.com/kamel-usp/jbcs2025`.

This research resulted in three publications. The first introduced a benchmark corpus of 385 essays across 38 different topics (also called prompts) with expert annotations from two independent graders, establishing baseline performance with encoder models (Silveira et al., 2024). The second investigated the robustness of transformer-based scorers against adversarial attacks, revealing vulnerabilities in both encoder and decoder architectures (Silveira et al., 2025). The third provided an extensive empirical analysis comparing 15 models across three paradigms, achieving state-of-the-art results with trait-specific QWK scores ranging from 0.60 to 0.73 (Barbosa et al., 2025).

The remainder of this extended abstract is organized as follows. Section 2 presents the conceptual framework including ENEM traits and information paradigms. Section 3 describes the methodology and presents experimental results. Section 4 concludes with practical implications and future directions. The full thesis is available at: `https://tinyurl.com/mvk34d98`

2 Concepts & Framework

The ENEM essay task assesses five traits that span different dimensions of writing quality. Table 1 summarizes these traits according to the official candidate guidelines. Traits C1 (Fluency) and C4 (Cohesion) evaluate surface-level linguistic features: grammar, spelling, punctuation, and the use of cohesive devices. These mechanical traits can be assessed with limited contextual information, as they depend primarily on the essay text itself. In contrast, C2 (Writing Style), C3 (Argumentation), and C5 (Persuasion) might require understanding the relationship between the essay and its prompt. Evaluating whether a student adequately addresses the topic, constructs relevant arguments, or proposes a viable intervention demands access to the prompt and supporting materials that define the task.

This work investigated cross-prompt, trait-specific scoring. In such a setting, models must evaluate essays on prompts not seen during training. Each trait is scored on a six-point ordinal scale $\{0, 40, 80, 120, 160, 200\}$, with the holistic (overall) score computed as the sum across all five traits (0–1000). The cross-prompt setting tests whether models learn transferable evaluation principles rather than prompt-specific patterns, reflecting realistic deployment scenarios where systems must generalize to new essay topics.

Two experimental dimensions systematically vary the information available to scoring models. The first dimension concerns *information access*: *prompt-blind* models receive only the essay text, while *prompt-aware* models receive the essay together with the prompt and supporting materials. This distinction is critical because mechanical traits (C1, C4) can theoretically be evaluated without prompt access, whereas argumentative traits (C2,

Code	Trait	Label	Description
C1	Trait 1	Fluency	Demonstrate command of the formal written modality of the Portuguese language.
C2	Trait 2	Writing Style	Understand the writing prompt and apply concepts from various fields of knowledge to develop the topic, within the structural limits of the argumentative-essay prose format.
C3	Trait 3	Argumentation and Relevance to Prompt	Select, relate, organize, and interpret information, facts, opinions, and arguments in defense of a point of view.
C4	Trait 4	Cohesion	Demonstrate knowledge of the linguistic mechanisms necessary for building argumentation.
C5	Trait 5	Persuasion/Intervention Proposal	Develop an intervention proposal for the issue addressed, while respecting human rights.

Table 1: Descriptions of the five ENEM essay scoring traits.

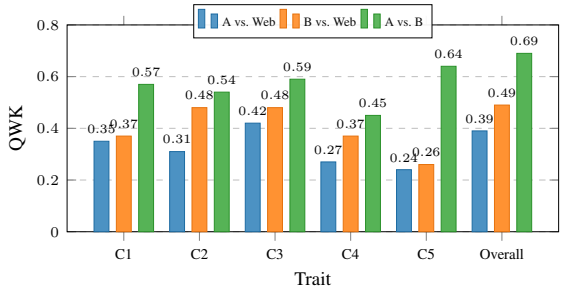


Figure 1: Human inter-rater agreement (QWK) across ENEM traits. Agreement between expert graders (A vs. B) ranged from 0.45 to 0.69 QWK.

C3, and C5) require understanding the task context. The second dimension concerns *rubric conditioning*, applicable only to decoder models operating in a zero-shot setting. Three rubric strategies were derived from official ENEM materials: Student guidelines (high-level descriptions provided to test-takers), Grader guidelines (detailed criteria from the official handbook), and Mixed guidelines (student-level descriptions with grader-level scoring rubrics).

Performance is measured using Quadratic Weighted Kappa (QWK), which quantifies ordinal agreement while penalizing predictions proportionally to their distance from ground truth (Cohen, 1960; de la Torre et al., 2018; Williamson et al., 2012; Ramnarain-Seetohul et al., 2022; Doewes et al., 2023). Standard bands classify QWK as poor (< 0.40), fair-to-good (0.40–0.74), or excellent (≥ 0.75) (Burrows et al., 2015; Fleiss et al., 2003), though these thresholds are not universally reliable (Bakeman et al., 1997). F1 Macro and F1 Weighted complement QWK by capturing class-level accuracy (Mello et al., 2024). As shown in Figure 1, human inter-rater agreement ranged from 0.45 to 0.69 QWK across traits, establishing a practical ceiling for automated systems.

3 Results and Discussion

Experiments use the dataset introduced by Silveira et al. (2024), comprising 385 essays across 38 prompts with independent annotations from two expert graders. When graders disagree, any prediction necessarily conflicts with at least one reference. Oracle baselines contextualize this ceiling: *Mean-Grade* (graders' rounded arithmetic mean) serves as the upper bound, while R_0 (most frequent training score) provides a lower bound.

The investigation systematically compared 15 models spanning three paradigms: *feature-based* methods using 72 linguistic features with Linear Regression and Random Forest classifiers; *encoder-only* transformers (109M–1.5B parameters) including BERTimbau, Albertina, and multilingual BERT; and *decoder architectures* divided into fine-tuned small language models (2.4B–14.7B parameters) such as Tucano, Phi-3, Llama3, and Phi-4, and zero-shot learners including GPT-4o, Sabiá3, and DeepSeek-R1 (up to 671B parameters). Table 2 summarizes model characteristics.

Category	Models	Params	Training
Feature-based	LR, RF	—	Full
Encoder-only	5 models	109M–1.5B	Full FT
Decoder (SLM)	4 models	2.4B–14.7B	LoRA
Decoder (ZSL)	3 models	up to 671B	Zero-shot

Table 2: Model categories evaluated. LR: Linear Regression; RF: Random Forest; FT: Fine-tuning; SLM: Small Language Models; ZSL: Zero-Shot Learners.

For decoder models, Figure 2 illustrates the prompt engineering framework exploring three dimensions: guideline source (Student, Grader, or Mixed), context inclusion (prompt-blind vs. prompt-aware), and response structuring via Chain-of-Thought (Wei et al., 2022) reasoning. This design yields six experimental conditions per trait for zero-shot evaluation.

Table 3 presents the best-performing configuration for each model class across all five traits, alongside feature-based baselines and oracle upper bounds. A detailed analysis of these interactions across all model configurations is provided in Barbosa et al. (2025).

Several key findings emerge from these results. Firstly, feature-based classifiers consistently under-

	C1: Fluency			C2: Writing Style			C3: Argument			C4: Cohesion			C5: Persuasion		
Model	M	W	Q	M	W	Q	M	W	Q	M	W	Q	M	W	Q
R_0 (Baseline)	.12	.20	.00	.11	.20	.00	.09	.17	.00	.14	.39	.00	.05	.05	.00
Linear Regressor	.41	.53	.36	.13	.23	.32	.17	.27	.26	.30	.57	.45	.15	.17	.03
Random Forest	.32	.56	.41	.14	.22	.22	.21	.29	.35	.35	.64	.48	.11	.15	.12
Best Encoder	.55	.71	.68	.33	.43	.32	.25	.36	.29	.48	.61	.60	.29	.37	.63
Best SLM	.52	.64	.67	.42	.52	.60	.37	.38	.57	.37	.58	.55	.44	.49	.59
Best ZSL	.35	.66	.69	.32	.43	.53	.44	.47	.73	.37	.60	.56	.37	.43	.60
Best	**.55**	**.71**	**.69**	**.42**	**.52**	**.60**	**.44**	**.47**	**.73**	**.48**	**.64**	**.60**	**.44**	**.49**	**.63**
MeanGrade (Oracle)	.71	.82	.85	.55	.71	.89	.62	.65	.88	.66	.83	.81	.68	.68	.90

Table 3: Test-set performance across model classes. M: macro F1, W: weighted F1, Q: QWK. SLM: Small Language Models (fine-tuned). ZSL: Zero-Shot Learners. MeanGrade represents the oracle upper bound.

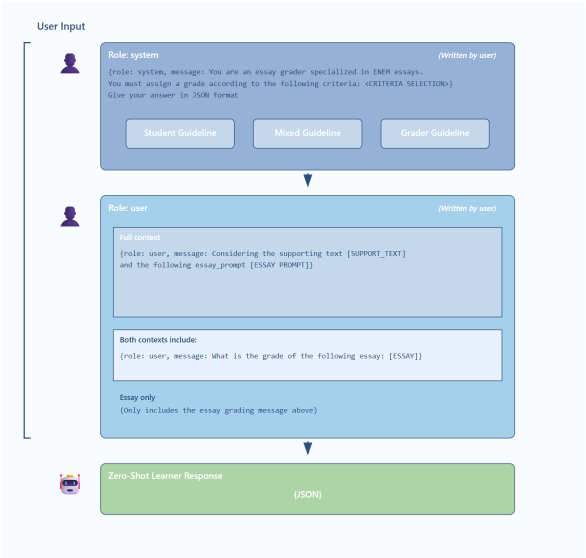


Figure 2: Prompt engineering framework for zero-shot essay scoring.

perform relative to neural approaches, with particularly poor results on Persuasion (C5). Secondly, no single architecture dominates across all traits: encoder models achieve the highest or comparable performance for Cohesion (C4: QWK 0.60) and Persuasion (C5: QWK 0.63); zero-shot learners excel at Argumentation (C3: QWK 0.73), likely due to emergent reasoning capabilities; and fine-tuned SLMs achieve the best results for Writing Style (C2: QWK 0.60). All three architectures achieve comparable results for Fluency (C1: QWK ranging from 0.67–0.69).

The impact of information access varies substantially across traits. For mechanical traits (C1, C4), additional context provides minimal benefit or even degrades performance, consistent with the "lost-in-the-middle" phenomenon where models struggle to retrieve information from extended sequences (Liu et al., 2024). In contrast, argumentative traits (C2, C3) show substantial gains when models receive the essay prompt and supporting materials.

Prompt engineering proves crucial for zero-shot performance. Notably, Sabiá3 (Abonizio et al., 2025), the Portuguese-specific model, outperforms multilingual alternatives only on Fluency (C1), suggesting that monolingual pretraining benefits surface-level linguistic assessment but confers minimal advantage for reasoning-intensive traits. A detailed analysis of model variants, context effects, and prompt engineering strategies is provided in Barbosa et al. (2025).

Comparison with oracle bounds reveals varying improvement potential across traits. Argumentation (C3: gap of 0.15) and Fluency (C1: gap of 0.16) approach human inter-rater agreement levels, while Writing Style (C2: gap of 0.29), Persuasion (C5: gap of 0.27), and Cohesion (C4: gap of 0.21) exhibit substantial headroom for advancement.

4 Conclusion

This dissertation explored how well automated systems can assess ENEM essays across five distinct traits. By systematically comparing 15 models under varying information access paradigms and rubric conditioning strategies, the research demonstrates that model performance varies significantly depending on which aspect of writing quality is being evaluated and what information is available to the model.

The findings demonstrate a striking pattern: models can almost reach human inter-rater agreement when evaluating mechanical aspects of writing, such as Fluency and Cohesion, as these features manifest as identifiable patterns within the text. However, significant disparities remain for traits that require deeper semantic understanding, including Writing Style, Argumentation, and Persuasion. This divergence indicates that current approaches are proficient at detecting linguistic indi-

cators but face challenges with tasks necessitating genuine comprehension of meaning and rhetorical effectiveness. Furthermore, no single configuration or architecture optimizes performance across all traits, which accounts for the consistent underperformance of unified approaches compared to specialized strategies.

Revisiting the research question, the evidence indicates that current AES systems occupy a middle ground between pattern matching and authentic evaluation. Evaluating whether arguments adequately respond to prompts, whether stylistic choices serve communicative goals, or whether intervention proposals present viable solutions requires reasoning capabilities beyond current systems. While these models detect surface-level quality markers, they do not engage with textual meaning the way humans do. Whether AES systems achieve genuine evaluation or merely sophisticated pattern matching remains an open challenge.

Several directions remain for future investigation. First, performance gaps for Writing Style (C2), Argumentation (C3), and Persuasion (C5) suggest these traits should receive focused attention in subsequent research. Second, encoder-only models with extended context windows, such as ModernBERT (Warner et al., 2024), could address current limitations of encoder-based approaches, though no such models existed for Portuguese at the time of this research. Third, exploring few-shot learning strategies may bridge the gap toward human inter-rater agreement levels beyond what context-limited zero-shot approaches achieve. Fourth, leveraging large language models for synthetic data generation could expand dataset size while preserving annotation consistency, enabling better performance for fine-tuned smaller models. Fifth, incorporating tool usage capabilities (Schick et al., 2023; Yao et al., 2023) could enable models to selectively retrieve trait guidelines or supporting materials only when needed, potentially mitigating the lost-in-the-middle effects observed with full-context inputs.

Rather than pursuing full automation, the field should embrace human-AI collaboration: automated systems handle mechanical evaluation while human graders focus on semantic dimensions where their understanding remains irreplaceable. The question is not whether machines can replace teachers, but how human and artificial intelligence can deliver more frequent, detailed feedback than either could provide alone.

4.1 Publications

This dissertation resulted in three peer-reviewed publications, summarized in Table 4.

Publication	Type	Year	Role
PROPOR (Silveira et al., 2024)	Conference	2024	Co-author
BRACIS (Silveira et al., 2025)	Conference	2025	Co-author
JBCS (Barbosa et al., 2025)	Journal	2025	First author

Table 4: Publications from this dissertation.

Acknowledgements

This work was partially supported by the São Paulo Research Agency (FAPESP) Grant no. 2022/02937-9, CNPq Grant no. 305136/2022-4 and CAPES Finance Code 001.

References

Hugo Abonizio, Thales Sales Almeida, Thiago Laitz, Roseval Malaquias Junior, Giovana Kerche Bonás, Rodrigo Nogueira, and Ramon Pires. 2025. Sabiá-3 technical report. *Preprint*, arXiv:2410.12049.

Dimitrios Alikaniotis, Helen Yannakoudakis, and Marek Rei. 2016. Automatic text scoring using neural networks. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 715–25.

Evelin Amorim and Adriano Veloso. 2017. A multi-aspect analysis of automatic essay scoring for Brazilian Portuguese. In *Proceedings of the Student Research Workshop at the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 94–102.

Yigal Attali. 2013. *Validity and Reliability of Automated Essay Scoring*. Routledge.

Yigal Attali and Jill Burstein. 2006. Automated essay scoring with e-rater v.2. *Journal of Technology, Learning, and Assessment*, 4(3).

R. Bakeman, D. McArthur, V. Quera, and B. F. Robinson. 1997. Detecting sequential patterns and determining their reliability with fallible observers. *Psychological Methods*, 2:357–370.

André Barbosa, Igor Cataneo Silveira, and Denis Deratani Mauá. 2025. An empirical analysis of large language models for automated cross-prompt essay trait scoring in brazilian portuguese. *Journal of the Brazilian Computer Society*, 31(1):857–870.

Steven Burrows, Iryna Gurevych, and Benno Stein. 2015. The eras and trends of automatic short answer grading. *International Journal of Artificial Intelligence in Education*, 25(1):60–117.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Jordi de la Torre, Domenec Puig, and Aida Valls. 2018. Weighted kappa loss function for multi-class classification of ordinal data in deep learning. *Pattern Recognition Letters*, pages 144–154. Machine Learning and Applications in Artificial Intelligence.

Afrizal Doewes, Nughthoh Arfawi Kurdhi, and Akrati Saxena. 2023. Evaluating quadratic weighted kappa as the standard performance metric for automated essay scoring. In *Proceedings of the 16th International Conference on Educational Data Mining*, pages 103–113. International Educational Data Mining Society.

Fei Dong, Yue Zhang, and Jie Yang. 2017. Attention-based recurrent convolutional neural network for automatic essay scoring. In *Proceedings of the 21st Conference on Computational Natural Language Learning*, pages 153–162.

Joseph L. Fleiss, Bruce Levin, and Myunghee Cho Park. 2003. *The Measurement of Interrater Agreement*, chapter 18. John Wiley & Sons, Ltd.

Erick Rocha Fonseca, Ivo Medeiros, Dayse Kamikawachi, and Alessandro Bokan. 2018. Automatically grading brazilian student essays. In *Proceedings of International Conference on Computational Processing of the Portuguese Language*, pages 170–179.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Watheq Ahmad Mansour, Salam Albatarni, Sohaila Eltanbouly, and Tamer Elsayed. 2024. Can large language models automatically score proficiency of written essays? In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2777–2786.

Jeziel Marinho, Rafael Anchiêta, and Raimundo Moura. 2021. Essay-br: a brazilian corpus of essays. In *Anais do III Dataset Showcase Workshop*, pages 53–64.

Rafael Ferreira Mello, Hilário Oliveira, Moésio Wenceslau, Hyan Batista, Thiago Cordeiro, Ig Ibert Bittencourt, and Seiji Isotanif. 2024. Propor'24 competition on automatic essay scoring of portuguese narrative essays. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese-Vol. 2*, pages 1–5.

Ellis B. Page. 1966. The imminence of... grading essays by computer. *The Phi Delta Kappan*, pages 238–243.

Vidasha Ramnarain-Seetohul, Vandana Bassoo, and Yasmine Rosunally. 2022. Similarity measures in automated essay scoring systems: A ten-year review. *Education and Information Technologies*, 27(4):5573–5604.

Pedro Uria Rodriguez, Amir Jafari, and Christopher M. Ormerod. 2019. Language models and automated essay scoring. *Preprint*, arXiv:1909.09482.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language models can teach themselves to use tools. *Preprint*, arXiv:2302.04761.

Igor Cataneo Silveira, André Barbosa, Daniel Silva Lopes da Costa, and Denis Deratani Mauá. 2025. Investigating universal adversarial attacks against transformers-based automatic essay scoring systems. In *Intelligent Systems*, pages 169–183, Cham. Springer Nature Switzerland.

Igor Cataneo Silveira, André Barbosa, and Denis Deratani Mauá. 2024. A new benchmark for automatic essay scoring in Portuguese. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 1*, pages 228–237.

Kaveh Taghipour and Hwee Tou Ng. 2016. A neural approach to automated essay scoring. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1882–1891, Austin, Texas. Association for Computational Linguistics.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Adams, Jeremy Howard, and Iacopo Poli. 2024. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. *Preprint*, arXiv:2412.13663.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

David M. Williamson, Xiaoming Xi, and F. Jay Breyer. 2012. A framework for evaluation and use of automated scoring. *Educational Measurement: Issues and Practice*, 31(1):2–13.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. React: Synergizing reasoning and acting in language models. *Preprint*, arXiv:2210.03629.

Evaluating FrameNet-Based Semantic Modeling for Gender-Based Violence Detection in Clinical Records

Lívia Dutra[1,2], Arthur Lorenzi[1,3], Frederico Belcavello[1], Ely Matos[1]
Marcelo Viridiano[1], Lorena Larré[1], Olívia Guaranha[3], Erik Santos[3]
Sofia Reinach[3], Pedro de Paula[3], Tiago Torrent[1,4]
[1]Federal University of Juiz de Fora (FrameNet Brasil)
[2]University of Gothenburg, [3]Vital Strategies Brasil
[4]Brazilian National Council for Scientific and Technological Development (CNPq)
livia.vicente.dutra@svenska.gu.se, {marcelo.viridiano, lorena.tasca}@estudante.ufjf.br
{alorenzi, oguaranha, esantos, pcbpaula, sreinach}@vitalstrategies.org
{fred.belcavello, ely.matos, tiago.torrent}@ufjf.br

Abstract

Gender-based violence (GBV) is a major public health issue, with the World Health Organization estimating that one in three women experiences physical or sexual violence by an intimate partner during her lifetime. In Brazil, although healthcare professionals are legally required to report such cases, underreporting remains significant due to difficulties in identifying abuse and limited integration between public information systems. This study investigates whether FrameNet-based semantic annotation of open-text fields in electronic medical records can support the identification of patterns of GBV. We compare the performance of an SVM classifier for GBV cases trained on (1) frame-annotated text, (2) annotated text combined with parameterized data, and (3) parameterized data alone. Quantitative and qualitative analyses show that models incorporating semantic annotation outperform categorical models, achieving over 0.3 improvement in F1 score and demonstrating that domain-specific semantic representations provide meaningful signals beyond structured demographic data. The findings support the hypothesis that semantic analysis of clinical narratives can enhance early identification strategies and support more informed public health interventions.

1 Introduction

The World Health Organization estimates that one in three women experiences physical or sexual violence by an intimate partner at some point in her life (WHO, 2024). Gender-based violence (GBV) is therefore not only a social issue, but a major public health concern (Garcia-Moreno and Watts, 2011; Sweet, 2014; Öhman et al., 2020). In Brazil, healthcare professionals are legally required to report cases of violence. Yet underreporting remains widespread. Either because victims are unable or unwilling to report their experiences or because signs of violence go unrecognized within routine medical encounters. Research suggests that many professionals struggle to identify signs of abuse, lack appropriate support tools, and work within fragmented information systems that do not communicate effectively (Kind et al., 2013; Garbin et al., 2015).

Brazilian public health systems collect large amounts of data on hospitalizations, mortality, medical records, and violence notifications. However, these systems are not fully integrated and lack a shared individual identifier (Guaranha et al., 2025). As a result, it is difficult to follow trajectories of risk over time or across institutions. Most of this information is stored in parameterized fields, which facilitate statistical analysis but capture only structured aspects of clinical encounters. Electronic medical records, however, also include open-text fields where healthcare professionals describe symptoms, circumstances, and patient histories in more detail. These narrative records often contain rich descriptions of situations that may signal risk of violence, but they are rarely consider for analysis due to its complexity.

The research reported in this paper explores whether semantic analysis of clinical narratives can help identify potential cases of GBV earlier and more reliably. The underlying assumption is that linguistic patterns embedded in medical records may reveal indicators of violence that are not captured by structured data alone. In particular, we investigate the contribution of FrameNet-based annotation to identifying possible patterns of violence within routine health data.[1] To achieve that, three experimental setups are compared using a SVM classifier : (1) a model trained on manually and automatically frame-annotated open-text data; (2)

[1]This study is based on the first authors' master's thesis and has been presented as part of a book chapter to lustrate the social applicability of FrameNet (Gamonal et al.).

Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)
Vol. 2, pages 49–57

a model trained on annotated open-text combined with parameterized data; and (3) a model trained exclusively on parameterized information. Model performance is assessed using precision, recall, and F1-score, along with qualitative analysis of possible semantic patterns.

The results show that models incorporating FrameNet-based semantic information outperform those relying solely on structured data, achieving an F1-score of 0.772, compared to 0.461 for the model trained exclusively on parameterized data. This result, backed by the qualitative analysis of the findings, suggests that semantic analysis of clinical narratives can provide meaningful support for the identification of gender-based violence in primary healthcare settings.

2 Frame-Based Models of Linguistic Cognition

FrameNet is a corpus-based computational lexical database grounded in Frame Semantics, a theory within Cognitive Linguistics proposed by Fillmore (1982). In Frame Semantics, word meaning is not treated as self-contained. Instead, meaning is understood through mental representations, called *frames*, which capture shared knowledge about recurrent situations, the participants involved in them, and the relations between those participants. A *frame* can therefore be seen as a structured background against which individual words are interpreted. Understanding a lexical item presupposes familiarity with this wider conceptual structure, since the meaning of any single element depends on how it fits into the scenario as a whole.

For instance, consider the lexical unit *diagnose.v*, which evokes the `Diagnosing` frame — Figure 1. The verb does not simply refer to an action. Rather, it presupposes a healthcare context in which several roles are necessarily present. At a minimum, there must be a healthcare professional who makes the diagnosis and a patient whose condition is being evaluated, defined as Frame Elements in the theory. Without these participants, the situation would be difficult to interpret as a diagnosis. The frame also allows for additional elements, such as the method in which the diagnosis was performed or the time and place that it happened. Thus, when the lexical unit *diagnose.v* appears in a clinical record, it activates a rich scenario of healthcare assessment.

FrameNet was then created in 1997 to implement this theoretical framework in a systematic,

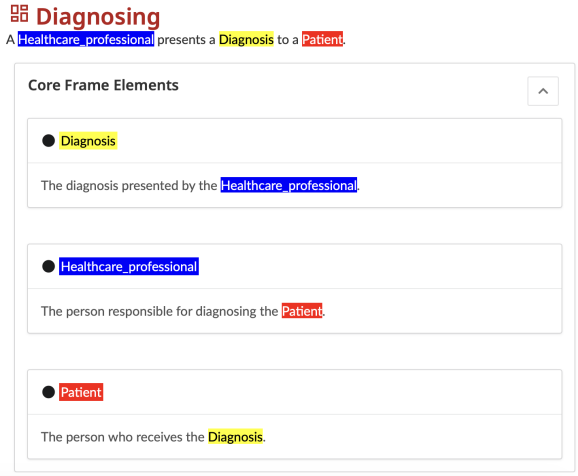


Figure 1: The `Diagnosing` frame.

corpus-based resource for English. The model has since been extended to several languages, including Brazilian Portuguese through FrameNet Brasil (FN-Br) (Torrent et al., 2022). Using its annotation tool (Torrent et al., 2024), FN-Br is able to link lexical units to the frames they evoke and annotate the semantic roles associated with those frames using authentic language data. In doing so, it is possible to gather insights and capture semantic representations that are both cognitively motivated and computationally interpretable for broader and more specific domains, as is the case of the current study, which focuses on semantic representations of healthcare and violence narratives. By modeling these domains explicitly and annotating relevant corpora, it becomes possible to identify recurring semantic configurations and relational patterns that may not be immediately visible in surface-level, as it is discussed next.

3 Corpora and Methods

This section presents the corpora and methods used in this study. The methodology involved modeling specific domains, manual and automatic annotation of data, and the design of GBV identification models and their quantitative and qualitative evaluation, as described next.

3.1 Corpora

The corpora used in this study were collected from health records produced in Recife, the capital of Pernambuco, Brazil, in collaboration with the Municipal Health Department. Data come primarily from two public health information systems: SINAN (Notifiable Diseases Information System)

and e-SUS AB (Primary Healthcare e-Medical Records System). In addition, definitions from the International Classification of Diseases (ICD) were incorporated to interpret diagnostic codes appearing in medical records. Also, causes of death were extracted from the Mortality Information System (SIM) as an indicator of true positive cases of GBV, in cases of violent deaths. The dataset is extensive, comprising more than three million records in e-SUS and more than 13,000 records in SINAN.

Regarding the systems structure, SINAN contains both parameterized fields and one open-text field ("observation") describing episodes of violence. For this study, only two components were used: (i) the parameterized indicator of a positive violence case and (ii) the corresponding open-text description. The e-SUS AB system records primary healthcare data through a combination of parameterized fields (e.g., age group, race, gender identity) and open-text fields following the SOAP model (Subjective, Objective, Assessment, Plan), along with additional narrative fields such as reason for referral, complements, and observations. Because diagnostic codes often appear without textual definitions in open-text fields, ICD descriptions were linked to the e-SUS records to enrich their semantic interpretability.

Given the presence of highly sensitive personal information in the corpora, strict ethical and legal safeguards were implemented as well as the anonymization of the data. The anonymization process combined automatic (NER models (Souza et al., 2019; Guillou, 2021; Cunha and Ramalho, 2022), fuzzy search of local place names, and regular expressions), semi-automatic (frequency-based detection of potential names), and manual verification methods to ensure the removal of Personally Identifiable Information (PII). Once this process was completed, anonymized samples of the data were used to model specific domains and for frame-based annotation.

3.2 Domain-Specific Modeling

As a means to fully capture the narratives in the open-fields of the public information systems, the specific semantic domains of Healthcare and Violence were modeled in FN-Br. This process entails the structuring of a cognitive representation that connects essential concepts of a given topic. This process consists of a twelve-step methodology, involving not only corpora collection and anonymization, but also lexical expansion, the modeling of new frames and relations between frames and lexical units – named Ternary Qualia Relations(Torrent et al., 2024) — and the lexicographic annotation of the corpora (Ruppenhofer et al., 2016), as described in Dutra et al. (2023) and Larré and Torrent (2024). This process was carried out by a group of ten researchers and resulted in 35 frames and 2,776 lexical units for the Healthcare domain along with 48 frames and 1,774 lexical units for the Violence domain.

3.3 Annotation

The annotation process was carried out in two phases: human and automatic. Human annotation followed the FrameNet methodology (Ruppenhofer et al., 2016) and was conducted using a sample of anonymized corpora with two aims. First, as the final stage of domain-specific modeling, annotation served the purpose of validating the domains; second, it was used to compose a dataset to train an automatic semantic labeler to be used in the entirety of the corpora. A total of seven trained annotators were part of the human annotation effort, which was carried out in a mirrored version of the FN-Br annotation tool (Torrent et al., 2024) with gated access to the data – so as to add one more layer of protection to the data being handled. The process focused on semantic annotation and consisted of selecting the frame evoked by each lexical unit in the sample sentences, which, then, generated an Annotation Set that allowed for the tagging of the frame elements represented in that narrative. The final number of annotated sentences was 2,352, resulting in over 14,600 Annotation Sets.

Automatic semantic labeling was performed using a newly trained version of LOME (Xia et al., 2021), a multilingual information extraction system that integrates a FrameNet parser within a pipeline based on XLM-RoBERTa, a BIO tagger and a Typer. For this study, LOME was trained in FrameNet Brasil's annotated data for English and Portuguese, including data from the Violence and Healthcare domains, expanding the training data used by Xia et al. (2021). The newly trained model had a micro-F1 of 50.68, slightly less than the original implementation (F1 = 56.34). Because this new instance was trained to deal with more challenging data, given its specificity, the performance is satisfactory. After training, the model was used to automatically annotate frames and frame frame elements in open-text sentences from e-SUS AB,

SINAN, and ICD records.

3.4 GBV Identification Models

As stated previously, this study focuses on evaluating the use of FrameNet-based semantic annotation to identify GBV cases and patterns in open-text medical records. In this sense, a model was developed to integrate FrameNet-annotated data and enable an assessment of feature importance for distinguishing violence from non-violence in e-SUS records.

Thus, as a means of accomplishing that, three experimental setups were conducted. All experiments used a linear Support Vector Machine (SVM). This choice was motivated by its interpretability, suitability for high-dimensional data, and consistency with the original project design. The three experimental setups used the same subset of e-SUS records, originally categorized based on ICD codes and links to SINAN notifications and SIM records. There were four labels:

- Violence: records with an ICD code for aggression or within two days of a SINAN notification or SIM record with the same code;
- Non-violence: ICD codes that have a small probability of being associated with violence, e.g. COVID-19 and some congenital malformations;
- Likely Violence: any record within 30 days of a notification of violence that does not have an ICD code for aggression;
- Unknown: any record that does not fall into one of the previous categories.

For this study, only violence and non-violence records were used, resulting in 801 cases (634 non-violence; 167 violence). Non-violence cases were undersampled to reduce class imbalance. Additionally, two specialists reviewed 100 "likely violence" cases, who reclassified them as 17 violence and 83 non-violence, increasing the complexity of the dataset by including more ambiguous cases.

The three experimental setups designed were:

1. **Semantic Model:** In the first setup, only the LOME annotated open-text fields were considered. As LOME identifies frame targets but not lexical units (LUs), an additional procedure to extract the LUs associated with each annotated span was also applied. This step increased the granularity of the representation, allowing the model to capture not only frames and frame elements, but also specific lexical choices. From these annotations, feature vectors were constructed based on the frequency of frames, frame elements, and lexical units, as well as the co-occurrence of frame elements across frames. Co-occurrences were considered only when at least one of the frames belonged to the Healthcare or Violence domains. The Ternary Qualia Relations between lexical units were also incorporated, assigning them a small weight to enrich the semantic connections without overwhelming the representation. To reduce sparsity, the least frequent features were removed (frames with fewer than 50 occurrences and LUs with fewer than 25). The resulting vectors were weighted using TF-IDF and L1-normalized. Given the high dimensionality of the semantic representation (15,456 features), Principal Component Analysis (PCA), was applied to reduce it to 2,000 components while preserving 94.8% of the variance. These components served as input to the classifier.

2. **Mixed Model:** The second setup considered annotated open-text fields and selected annotated parameterized fields. This experiment followed the same pipeline as the first, but additionally incorporated structured parameterized fields into the semantic representation. Categorical variables — such as race, gender identity, sexual orientation, prosthesis need, and age group — were mapped to corresponding lexical units and frames. After TF-IDF weighting, the feature space comprised 15,478 dimensions. PCA again reduced this to 2,000 components, preserving 93.5% of the variance. This combined representation was used as the input to the model.

3. **Demographic Model:** The third setup excluded semantic annotation and relied exclusively on parameterized data. The features included demographic, clinical, and administrative variables such as race, age, ICD codes, marital status, education level, unit location, and referral timing. Categorical variables were transformed using One-Hot Encoding, expanding the original 20 structured variables to 142 binary features. These features were

used directly as input to the classifier.

3.5 Evaluation

After training, the model was evaluated both quantitatively and qualitatively. While the quantitative evaluation provided numerical evidence of model performance, the qualitative analysis aimed to interpret and understand what the models were learning and how FrameNet annotation contributed to GBV identification.

Quantitative (SVM) Evaluation To compare the three experimental setups and assess the relevance of different data sources for GBV identification, the performance of the models was evaluated using five-fold cross-validation. In this procedure, the dataset was divided into five subsets: in each iteration, four subsets were used for training and one for testing, rotating until all subsets had served as the test set. Performance was assessed using precision, recall and F1 score, and the final results correspond to the average across the five folds.

Qualitative Evaluation This process involved two complementary lines of analysis: first, feature importance scores were extracted from the best-performing semantic model and the demographic model; second, the most frequently evoked frames and lexical units in the annotated e-SUS records of confirmed victims were also examined.

1. **Model features:** The 35 most relevant features for both the semantic and the demographic models were analyzed based on their contribution to the classification of the cases[2]. This allowed us to assess the explanatory power of parameterized fields versus semantic features and to identify key frames, frame elements, and lexical units associated with GBV cases.

2. **Frame and lexical pattern analysis:** Next, the frame activation patterns were analyzed in confirmed GBV cases by examining:

    - the 15 most frequently evoked frames in both domains — Healthcare and Violence;
    - the 20 most frequent LUs per domain;
    - the 30 most frequent LUs evoking the `Health_conditions` frame, to explore possible links between health conditions and violence.

    This analysis allowed a better understanding of patterns that could be linked to GBV and pointed towards future investigation.

[2]At this point, it is not possible to identify for which of the classes each feature was most relevant, only that they were relevant for the model's decision.

4 Results and Discussion

In this section, we present the results of the evaluation conducted on the model setups and discuss their implications to the use of frame-based representations to the identification of GBV in e-medical records.

4.1 Quantitative Evaluation: SVM Models

As shown in Table 1, the semantic model that relies solely on open-text fields consistently produced the strongest results. It obtained the highest recall, indicating an effective identification of positive cases. Precision was lower, but this trade-off is acceptable in the context of the study, once the main concern is avoiding false negative cases of violence in a setting where underreporting is an issue. The F1-score reflects this balance. Furthermore, recall values showed little variation across cross-validation splits, while precision and F1 varied more substantially, suggesting that some data splits were more challenging for the model than others.

Adding parameterized data to the semantic model did not lead to a meaningful improvement in the second experiment. Although a small increase in precision was observed, this was achieved without gains in overall performance and is not particularly advantageous for the task at hand, as it increases the likelihood of false negatives. These results reinforce the idea that semantic information extracted from textual descriptions is more important in identifying cases of violence than structured demographic attributes.

Finally, the contrast with the demographic model is clear. Although recall values were relatively high, precision was extremely low, indicating that many cases were incorrectly classified as positive. This imbalance makes the recall results difficult to interpret with confidence. Moreover, recall varied widely across different splits, resulting in consistently low F1-scores. Taken together, these results show that parameterized data alone are not sufficient for reliable case identification, even if they may be useful for descriptive analyzes.

Overall, the results support the initial hypothesis that FrameNet-based semantic analysis

Model	F1	Recall	Precision
Semantic	0.772 (0.113)	0.838 (0.071)	0.756 (0.190)
Mixed	0.771 (0.114)	0.832 (0.078)	0.759 (0.189)
Demographic	0.461 (0.089)	0.701 (0.173)	0.345 (0.057)

Table 1: Model performance comparison

contributes meaningfully to the identification of gender-based violence cases in electronic medical records. Next, these findings are complemented with a qualitative analysis of the patterns identified by this approach.

4.2 Qualitative Evaluation: Models and Domains

Demographic Model The list of the 35 most relevant features for the demographic model consists of the fill-in options associated with the parameterized fields, as shown in Figure 2. These features provide limited insight, as — at this stage — it is not possible to determine whether they contributed to the classification of violence or non-violence cases.

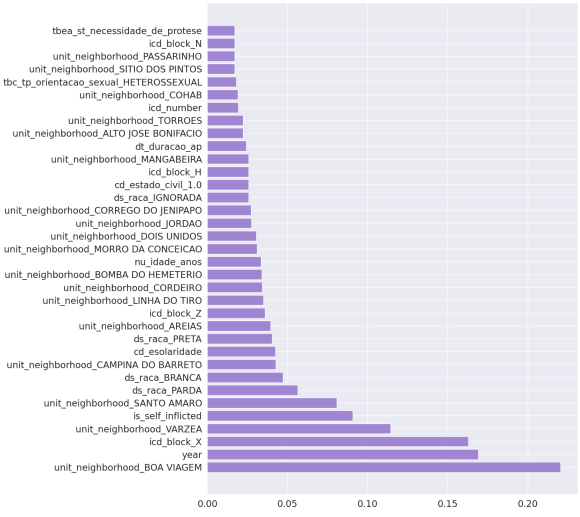


Figure 2: Most relevant features of the Demographic Model

The feature of greatest relevance is the neighborhood in which a health unit is located. In fact, 19 of the 35 selected features are related to the location of healthcare facilities. The location of a health unit does not have a clear explanatory link to the occurrence of violence, particularly since patients may seek care in different facilities, making geographic information a weak and potentially misleading indicator.

Race also appears as a highly relevant feature. Of the six possible values, four — *branca* ('white'), *parda* ('brown'), *preta*('black'), and *ignorada* ('ignored') — were selected, with three among the top ten. This pattern can be influenced by inconsistent field completion, which can amplify the weight of filled values. As a result, the model emphasizes individual characteristics rather than contextual information, reinforcing stereotypes this work aims to avoid.

Despite these limitations, some relevant features are associated with the type of care sought, notably ICD blocks X, Z, H, and N. These correspond to external causes of morbidity and mortality, factors influencing health status and contact with health services, diseases of the eye and adnexa, and diseases of the genitourinary system. Although not predominant, the presence of blocks N and Z is consistent with the patterns identified by the semantic model, as block N may relate to sexual violence and block Z often reflects scenarios requiring follow-up care, such as prenatal monitoring, that are aligned with the findings in the semantic setup.

Thus, this qualitative analysis reinforces the quantitative findings. Parameterized data leads to a model focused primarily on individual attributes, with limited attention to the clinical context. Meaningful interpretation of care-related features was only possible through a comparison with the semantic model, further supporting the use of FrameNet-based semantic analysis for open-text fields. The next section examines the most relevant features of the semantic model that had the best performance in the quantitative analysis.

Semantic Model Figure 3 shows the features that most influenced the semantic model, including frames, frame elements, co-occurrences between frame elements, and lexical units. At this stage, the analysis is exploratory and the discussion focuses on recurring patterns rather than on a detailed interpretation.

Relevant features are not restricted to the Healthcare and Violence domains. Among the most influential features are general vocabulary frames, most notably `Personal_relationships`. This result is unsurprising, given that many cases of gender-based violence involve individuals who are related in some way, which may reflect indirect references to aggressors in the records. Another generic frame

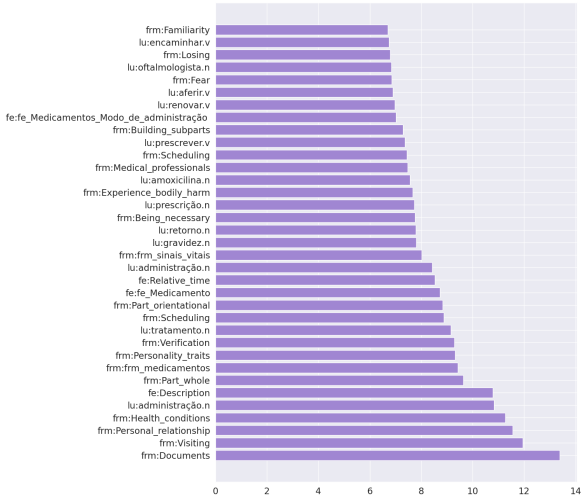


Figure 3: Most Relevant Features of the Semantic Model

that appears prominently is `Fear`. Although it is not part of the Violence domain, its relevance suggests that emotional states expressed in the text may contribute to identifying situations associated with violence.

Healthcare-related frames are more frequent than violence-related ones. Only the `Experience_bodily_harm` frame appears as a relevant frame from the Violence domain. However, the presence of this frame supports the idea that violence-related patterns can be inferred from clinical information. This is particularly relevant for primary care settings, where early identification is critical.

Within the Healthcare domain, `Health_Conditions` ranks among the most relevant frames, which is expected given the nature of the data. Frames related to `Medicines` also appear repeatedly, including more than one of their frame elements. In addition, several lexical units associated with the `Health_Intervention` frame appear as relevant for the model's decisions. Together, these results indicate that routine clinical actions and treatment-related information play an important role in the identification process.

The semantic features identified in the model provide a starting point for understanding how patterns related to gender-based violence can appear in clinical narratives. Rather than examining all relevant patterns, the focus was on the most frequently evoked frames — and lexical units — annotated by LOME in the confirmed cases of GBV that are part of the Healthcare and Violence domains.

Within the Healthcare domain, `Health_conditions` is by far the most frequently evoked frame. This is expected given the nature of clinical records and it was also evident in the model's features' analysis. However, only a small number of its associated lexical units appeared among the most frequent terms, suggesting that not all health conditions carry the same analytical weight. One term that stands out is *gestante.n* ('pregnant'), while *gravidez.n* ('pregnancy') was also a relevant feature for the semantic model. The prominence of pregnancy-related terms raises an interpretive challenge: it is unclear whether this reflects increased vulnerability, patterns specific to the dataset, or a potential bias toward the specific gender healthcare event in focus.

The `Health_service` frame also plays an important role. Terms such as *encaminhar.v* ('referral') and *acompanhamento.n* ('follow-up') occur frequently and often refer to specialized care or ongoing treatment. When considered together with other clinical elements, these references may indirectly signal previous incidents. Similarly, frequent mentions of medical examinations and medications suggest that routine clinical procedures may carry contextual clues. Enriching these elements through semantic relations, such as ternary qualia relations, may allow deeper inferences about the underlying conditions and possible links to violence.

A closer look at the `Health_conditions` frame reveals two notable tendencies. First, pregnancy-related terms appeared with high frequency, as it was already pointed out. Second, mental health conditions — including depression, anxiety, and bipolar disorder — were highly represented. These patterns may reflect the psychological consequences of abuse, although they may also be influenced by broader gendered healthcare-seeking behaviors. In either case, they warrant further investigation.

In contrast, frames from the Violence domain appear less frequently, likely because, in comparison to health issues, explicit references to violence are less frequent in clinical records. Among them, the `Experience_bodily_harm` frame stands out and also contributed to the performance of the model. However, many of its associated terms, such as fall or trauma, are not inherently indicative of violence. More direct signals emerge in references to self-inflicted harm, including self-mutilation and suicide. Although these cases were not analyzed separately, their frequency suggests that self-directed

violence deserves closer attention in future work.

Sexual violence-related patterns are particularly salient. Terms associated with sexual acts, abuse, and related examinations appear consistently, indicating that sexual violence may be a significant factor motivating healthcare visits. References to sexually transmitted infections further reinforce this interpretation. These patterns suggest that even when violence is not explicitly documented, its consequences may be traceable through clinical descriptions.

Therefore, this qualitative analysis also shows that FrameNet-based semantic annotation makes it possible to uncover patterns that would likely remain invisible in structured data alone. Although the findings remain exploratory and require validation in collaboration with healthcare professionals, they support the broader hypothesis that semantic analysis of open-text medical records can contribute to the identification of gender-based violence in primary care settings.

5 Conclusion and Future Work

This study evaluated the use of FrameNet-based semantic annotation to identify Gender-Based Violence (GBV) cases and patterns in open-text fields of e-medical records. Our results show that models using semantic annotation of open-text fields outperform models relying solely on parameterized demographic data, achieving an F1 score 0.31 higher. The addition of parameterized fields to the semantic model provided minimal improvement, highlighting that open-text information carries richer and more relevant insights for detecting GBV. Qualitative analysis confirmed that relying only on parameterized data risks reinforcing stereotypes and provides limited information for pattern discovery, while semantic annotation enables the identification of meaningful patterns that can inform further investigation and policy intervention.

In general, these findings support the hypothesis that FrameNet-based semantic analysis is a valuable tool for identifying potential GBV cases, including those underreported. By revealing patterns in both reported and unreported cases, this approach can assist in early-warning systems and public policies, contributing to improved protection and intervention strategies. Finally, the experimental setups and qualitative assessments presented here provide a baseline for future research on leveraging linguistic analysis for public health surveillance, which is already in development. Progress has already been made on the more systematic identification of GVB patterns (Dutra et al., 2025), and three parallel lines of research are currently being carried out to explore the identification of new patterns. Two of them focus directly on violence-related cases, specifically self-harm and violence against LGBTQ+ individuals. The third broadens the scope of the semantic model beyond violence, with the aim of identifying patterns related to women's health and supporting early detection of potential cancer cases.

Ethics and Limitations

The study presented in this paper was part of a broader project and approved by the Research Ethics Committee of the Federal University of Goiás (CAAE:64733922.3.0000.5083; Approval number: 6.126.995). The research involved highly sensitive information from violence notifications and electronic medical records that could increase the risk for victims of violence. Thus, the research team has extensively studied this issue and consulted data protection specialists before pursuing any implementation of the methodology. To protect the information, all team members signed confidentiality agreements, the data was anonymized - as described-, and access was restricted to anonymized samples only. The methodology was developed to improve the use of health data in Brazil and address the underreporting of health-related events, using frame-based modeling, semantic parsing, and the identification of linguistic pattern in Brazilian Portuguese. Although it can be adjusted and expanded to other languages, it has not been extensively tested yet and may reflect biases specific to Brazilian Portuguese, which is a limitation.

Acknowledgments

This work was supported by the Patrick J. McGovern Foundation's acceleration program, the José Luiz Setúbal Foundation, and the Instituto Galo da Manhã. We also express our gratitude to our partners from the Recife Municipal Health Department — Luciana Caroline, Marcella Abath, Natalia Barros, and Yana Lopes — for their valuable collaboration and continuous support throughout this project. Tiago Torrent is a grantee of the Brazilian National Council for Scientific and Technological Development CNPq – grant 311241/2025-5).

References

Luís Filipe Cunha and José Carlos Ramalho. 2022. Ner in archival finding aids: Extended. *Machine Learning and Knowledge Extraction*, 4(1):42–65.

Lívia Dutra, Arthur Lorenzi, Laís Berno, Franciany Campos, Karoline Biscardi, Kenneth Brown, Marcelo Viridiano, Frederico Belcavello, Ely Matos, Olívia Guaranha, Erik Santos, Sofia Reinach, and Tiago Timponi Torrent. 2025. Frame semantic patterns for identifying underreporting of notifiable events in healthcare: The case of gender-based violence. *Preprint*, arXiv:2510.26969.

Lívia Dutra, Arthur Lorenzi, Lorena Larré, Frederico Belcavello, Ely Matos, Amanda Pestana, Kenneth Brown, Mariana Gonçalves, Victor Herbst, Sofia Reinach, Renato Teixeira, Pedro de Paula, Alessandra Pellini, Cibele Sequeira, Ester Sabino, Fábio Leal, Mônica Conde, Regina Grespan, and Tiago Torrent. 2023. Building a frame-semantic model of the healthcare domain: Towards the identification of gender-based violence in public health data. In *Anais do XIV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 338–346, Porto Alegre, RS, Brasil. SBC.

Charles J. Fillmore. 1982. Frame Semantics. In Linguistics Society of Korea, editor, *Linguistics in the morning calm*. Hanshin Publishing Co., Seoul, South Korea.

Maucha Andrade Gamonal, Lívia Vicente Dutra, Mariane de Carvalho Pinto, and Tiago Timponi Torrent. Pln e responsabilidade social: Aplicações da framenet-br. In H. M. Caseli and M. G. V. Nunes, editors, *Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português*, 4 edition, volume 3. BPLN. In press.

Cléa Adas Saliba Garbin, Isabella de Andrade Dias, Tânia Adas Saliba Rovida, and Artênio José Ísper Garbin. 2015. Desafios do profissional de saúde na notificação da violência: obrigatoriedade, efetivação e encaminhamento. *Revista Ciência & Saúde Coletiva*, 20(6):1879–1890.

Claudia Garcia-Moreno and Charlotte Watts. 2011. Violence against women: an urgent public health priority. *Bulletin of the World Health Organization*, 89:2–2.

Olívia LC Guaranha, Juliana Rocha Miranda, Fátima Marinho, Renato Teixeira, Erik Santos, Denise Guerra Wingerter, Paola da Costa Silva, Diana Paula de Souza Rego Pinto, Gleidson Paulino Vitório, Sofia Reinach, and 1 others. 2025. Data integration for the prevention of violence against girls and women in northeastern brazilintegración de datos para la prevención de la violencia contra niñas y mujeres en el nordeste de brasil. *Revista panamericana de salud publica= Pan American journal of public health*, 49:e66.

Pierre Guillou. 2021. NER-BERT-Base-Cased-pt-lenerbr: BERT-based NER model for Portuguese (legal domain). https://huggingface.co/pierreguillou/ner-bert-base-cased-pt-lenerbr. Accessed: 2023-08-30.

Luciana Kind, Maria de Lourdes Pereira Orsini, Valdênia Nepomuceno, Letícia Gonçalves, Gislaine Alves de Souza, and Monique Fernanda Félix Ferreira. 2013. Subnotificação e (in) visibilidade da violência contra mulheres na atenção primária à saúde. *Cadernos de Saúde Pública*, 29:1805–1815.

Lorena Larré and Tiago Torrent. 2024. Modelagem baseada em frames para identificação do léxico da violência de gênero. In *Anais do XV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 403–412, Porto Alegre, RS, Brasil. SBC.

Josef Ruppenhofer, Michael Ellsworth, Miriam Petruck, Christopher Johnson, and Jan Scheffczyk. 2016. *FrameNet II: Extended Theory and Practice*.

Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. 2019. Portuguese Named Entity Recognition using BERT-CRF. *arXiv preprint arXiv:1909.10649*.

Patricia L Sweet. 2014. Every bone of my body: Domestic violence and the diagnostic body. *Social Science & Medicine*, 122:44–52.

Tiago Timponi Torrent, Ely Edison da Silva Matos, Frederico Belcavello, Marcelo Viridiano, Maucha Andrade Gamonal, Alexandre Diniz da Costa, and Mateus Coutinho Marim. 2022. Representing Context in FrameNet: A Multidimensional, Multimodal Approach. *Frontiers in Psychology*, 13.

Tiago Timponi Torrent, Ely Edison da Silva Matos, Alexandre Diniz da Costa, Maucha Andrade Gamonal, Simone Peron-Corrêa, and Vanessa Maria Ramos Lopes Paiva. 2024. A flexible tool for a qualia-enriched FrameNet: the FrameNet Brasil WebTool. *Language Resources and Evaluation*, pages 1–29.

WHO. 2024. World Health Organization Health Topics: Violence Against Women. Accessed: October 6, 2025.

Patrick Xia, Guanghui Qin, Siddharth Vashishtha, Yunmo Chen, Tongfei Chen, Chandler May, Craig Harman, Kyle Rawlins, Aaron Steven White, and Benjamin Van Durme. 2021. LOME: Large ontology multilingual extraction. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 149–159, Online. Association for Computational Linguistics.

Ann Öhman, Marika Burman, Maria Carbin, and Kerstin Edin. 2020. The public health turn on violence against women: analysing swedish healthcare law, public health and gender-equality policies. *BMC Public Health*, 20:1–12.

Pretrained Neural Audio Models for Asthma Detection from Voice and Speech

Leticia Puttlitz Boll
Universidade de São Paulo (USP), Brazil
leticia.puttlitz@usp.br

Antonio Oss Boll
Universidade de São Paulo (USP), Brazil
aoboll@usp.br

Yan Anderson Pires de Oliveira
Universidade de São Paulo (USP), Brazil
yananderson@usp.br

Victor dos Santos Silva
Universidade de São Paulo (USP), Brazil
victorsantos@usp.br

Mariana Lopes Pestana
Universidade de São Paulo (USP), Brazil
marylopestana@usp.br

Celso Ricardo Fernandes de Carvalho
Universidade de São Paulo (USP), Brazil
cscarval@usp.br

Marcelo Matheus Gauy
Universidade Estadual Paulista (UNESP), Brazil
marcelo.gauy@unesp.br

Marcelo Finger
Universidade de São Paulo (USP), Brazil
mfinger@ime.usp.br

Abstract

Asthma is a chronic respiratory disease that affects breathing and may also influence speech and voice production. In this paper, we examine whether short mobile-recorded Brazilian Portuguese voice and speech audio contain cues that can be used to distinguish individuals with asthma from those without asthma. We approach this problem using transfer learning with pretrained neural audio models based on convolutional architectures trained on large-scale audio datasets (PANNs). We evaluate two recording types: sustained vowel phonation and read speech. Models are trained for a binary classification task and evaluated at both the segment level and the patient level. Read speech performs better than sustained vowels. The best configuration (CNN14 on speech) achieves 0.85 patient-level balanced accuracy (accuracy 0.85) with ROC-AUC 0.93 and PR-AUC 0.98, performing comparably to CNN10. Training from scratch performs worse than fine-tuning a pretrained model, showing that pretraining helps when data is limited. Performance also varies across age groups, suggesting demographic sensitivity. These findings support the feasibility of audio-based asthma classification from voice and speech and motivate further investigation of pretrained audio models in biomedical applications.

1 Introduction

Asthma is a chronic airway disease characterized by airflow limitation and symptoms such as wheezing, dyspnea, chest tightness, and cough, which may lead to exacerbations requiring hospitalization (Global Initiative for Asthma, 2024; Chipps et al., 2023; Fuhlbrigge et al., 2012). Because asthma affects breathing patterns, it can also introduce acoustic changes in voice and speech.

Advances in deep learning have enabled powerful models for audio processing, such as PANNs (Kong et al., 2020), which are pretrained on thousands of hours of audio and can achieve strong performance even with limited labeled data. This makes pretrained audio models a promising approach for asthma detection from short recordings. In this context, voice and speech data collected via mobile devices provide a non-invasive and low-cost data source that can be analyzed with machine learning to investigate clinically relevant acoustic patterns for respiratory disease classification.

In this work, we focus on the binary task of distinguishing *asthma* vs. *non-asthma* using short recordings of sustained vowels and read speech. We evaluate pretrained convolutional audio networks adapted to this task. We compare shallow and deeper architectures, pretrained versus scratch models, and analyze demographic subgroups to assess robustness. On read speech, the pretrained CNN14 achieves the best patient-level balanced accuracy (0.85) and accuracy (0.85), with ROC-AUC 0.93 and PR-AUC 0.98, although its performance is comparable to CNN10. The pretrained CNN10 shows slightly higher accuracy (0.86) but lower

Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)
Vol. 2, pages 58–67

balanced accuracy, and both outperform sustained vowels and scratch models.

2 Related Work

2.1 Pretrained Neural Audio Models

Pretrained neural audio models have increasingly been adopted in healthcare and related audio applications due to their ability to learn robust acoustic representations. Such models have demonstrated strong performance across diverse tasks, including sound event detection (Xu et al., 2023), emotion recognition (Gauy and Finger, 2022), and respiratory disease identification (Gauy et al., 2023; Matheus Gauy et al., 2026).

In the context of emotion recognition, Gauy and Finger (2022) showed that large-scale audio pretraining improves performance, allowing models to outperform baselines even with limited labeled data. Similar gains have been reported in voice-based neurological disorder detection, where pretrained convolutional neural networks operating on spectrograms of sustained vowel recordings outperformed models based on handcrafted acoustic features for Parkinson's disease classification (Rahmatallah et al., 2025).

Furthermore, results on the OPERA respiratory audio benchmark indicate that models pretrained on large and diverse datasets such as AudioSet consistently surpass both models trained from scratch and those pretrained exclusively on respiratory sounds, reinforcing the value of large-scale general-domain pretraining for medical audio analysis (Niizumi et al., 2025).

2.2 Machine Learning for Asthma Detection

Artificial intelligence and machine learning have been increasingly applied to asthma screening, phenotyping, and disease monitoring across a wide range of clinical and biomedical data modalities (Exarchos et al., 2020). Prior work includes asthma classification using machine-learning models trained on pulmonary function test results combined with clinical variables (Topalovic et al., 2017), as well as leveraging cough acoustics as a complementary audio biomarker (Alqudaihi et al., 2021). In addition to acoustic signals, other respiratory measurements have also been explored, including quantitative features derived from exhaled CO_2 waveforms (Singh et al., 2018) and breath-based biomarkers such as exhaled nitric oxide for diagnosis and severity monitoring (Yin et al., 2025). Beyond respiratory signals, asthma classification has further been investigated using molecular biomarkers, such as nasal gene-expression signatures (Pandey et al., 2018), and routine blood biomarkers modeled with machine-learning techniques (Zhan et al., 2020).

In a related line of work focusing on voice-based biomarkers for asthma, several studies have relied on sustained vowel recordings. An XGBoost-based classifier was proposed using handcrafted acoustic features extracted from sustained vowel /a:/ recordings (Lyu et al., 2025). Similarly, MeLoDicA (Looi et al., 2024) introduced a framework based on handcrafted spectral and temporal voice features, showing that sustained vowels achieved the highest performance among the evaluated audio types.

Beyond sustained vowels, asthma detection has also been explored using speech signals. Real-time asthma classification using speech and respiratory sounds has been investigated (Iqbal et al., 2022), and conventional classifiers such as GMMs and CNNs have been applied to MFCC features extracted from speech (Iqbal et al., 2024). In addition, machine-learning models have been trained on short Turkish phonetic utterances (Gezer et al., 2025).

Overall, these approaches rely predominantly on manually designed acoustic features and conventional machine-learning models, without leveraging large-scale pretrained audio representations or end-to-end learned embeddings.

2.3 Voice and Speech as Biomarkers in Respiratory Diseases

Recent reviews have highlighted the use of audio-based biomarkers for respiratory disease detection, including cough, lung sounds, and voice or speech signals (Kapetanidis et al., 2024). Beyond asthma, similar approaches have been applied to the identification of respiratory diseases. Pretrained audio models have been used to analyze speech and voice recordings from patients with respiratory conditions (Gauy et al., 2023; Matheus Gauy et al., 2026). Voice and speech have also been investigated as biomarkers for COVID-19 detection, where machine-learning models based on acoustic features have shown strong discriminative performance in identifying infected individuals (Verde et al., 2023; Dash et al., 2022). In addition, voice-based methods have been applied to chronic obstructive pulmonary disease (COPD), using embeddings from a wav2vec 2.0 model to classify disease

presence and severity from short voice recordings (Lee et al., 2025).

3 Data

We use a dataset of short voice recordings collected from adult speakers of Brazilian Portuguese performing two speaking tasks: *Speech* and *Vowel*. In the Speech task, participants read a short predefined sentence, while in the Vowel task they sustained the vowel /a:/ for as long as they could. These tasks were chosen to capture different aspects of speech production and respiratory control.

The recordings were collected using mobile devices and processed at a sampling rate of 16 kHz. Data were recorded both in hospital settings and in more uncontrolled environments using participants' personal devices. Because of this, the dataset includes variation in background noise, microphone quality, and recording conditions.

The dataset includes 549 Speech and 538 Vowel recordings, with asthma representing 79% of the samples. The average recording length is 7.5 seconds.

3.1 Clinical and demographic metadata

In addition to the audio recordings, clinical and demographic metadata are collected for each participant. These include age, sex, anthropometric measures (weight and height), vital signs, as well as information on comorbidities and smoking history. These variables provide important contextual information for the analysis and allow for the assessment of potential demographic biases in the data.

3.2 Demographic distribution

Table 1 summarizes the distribution of recordings by sex for both speaking tasks. The dataset is predominantly female in both cases.

Table 1: Sex distribution by speaking task (unique patients).

Sex	Speech	Vowel
Female	434	426
Male	115	112
Total	549	538

Table 2 reports the age distribution using four age bins: under 30 years, 30–45 years, 45–60 years, and over 60 years.

Table 2: Age distribution by speaking task (unique patients).

Age range (years)	Speech	Vowel
< 30	81	80
30–45	140	137
45–60	244	239
> 60	84	82
Total	549	538

4 Preprocessing

Using an energy-based trimming technique that eliminates low-energy regions in relation to the signal's peak, we eliminate silence at the start and finish of each audio file. The resulting waveform is then resampled to 16 kHz and peak-normalized by scaling each waveform to a fixed maximum absolute amplitude (Labied et al., 2022).

4.1 Dataset splitting

The dataset is divided into training, validation, and test sets using a stratified splitting strategy at the patient level, with proportions of 60%, 20%, and 20%, respectively. In order to guarantee that these demographic characteristics are evenly distributed throughout splits, stratification is carried out according to age group and sex (Xu and Goodacre, 2018).

4.2 Class balancing and data augmentation

The dataset is imbalanced with respect to the target classes. To mitigate this effect during model learning, we apply class balancing on the training set only. In our experiments, we use random oversampling, duplicating minority-class recordings until the class counts match those of the majority class.

In addition, to increase robustness to recording variability, we apply waveform perturbations during preprocessing for training examples selected as augmented duplicates. Specifically, one of the following transformations is sampled uniformly at random: additive Gaussian noise, random gain perturbation, pitch shifting, or time stretching (Wei et al., 2020). These operations are applied at the waveform level prior to feature extraction.

4.3 Temporal windowing

We use a sliding-window approach to create fixed-length segments because recordings vary in length. Following previous literature, each waveform is split into 4.0 s windows with a 2.0 s hop (Casanova

et al., 2021), producing a variable number of segments depending on the length of the recording. Recordings are zero-padded to 4.0 s if they are shorter than the window length. Although segments from the same recording may overlap, windowing is performed after splitting the dataset at the patient level. This ensures that segments from the same speaker are only present in one split, preventing similar audio samples from appearing in both training and test sets.

4.4 Noise injection

To further reduce sensitivity to background conditions, we inject environmental noise by mixing each window with a randomly selected noise sample drawn from a separate noise pool. This pool consists of recordings collected in the same hospital environments as the patient data. The noise is added to the signal with a randomly scaled amplitude after being trimmed to the same length as the window. The noise amplitude is randomly scaled to simulate different noise levels while ensuring that the added noise does not dominate or distort the speech signal.

By using this technique, the model is prevented from learning hospital background noise as a cue for asthma, for example, or from linking background noise with the target labels. The model is encouraged to concentrate on vocal and speech-related information rather than environmental artifacts by changing background conditions independently of the labels.

4.5 Time-frequency representations

For each window, the audio signal is converted into a log-Mel spectrogram, a time-frequency representation that summarizes how the signal energy is distributed over time and frequency. This representation is used as input to the neural models.

Figure 1 summarizes the full preprocessing pipeline used in our experiments, from waveform normalization and participant-level splitting to windowing, augmentation, noise mixing, and feature extraction.

5 Models

We use pre-trained convolutional neural networks for audio classification, specifically the CNN10 and CNN14 architectures, as proposed in the PANNs framework (Kong et al., 2020). These models were pre-trained on the AudioSet database (Gemmeke et al., 2017), which contains over two million labeled audio clips corresponding to more than 5,000 hours of audio, and are widely used in prior work as general-purpose audio feature extractors.

The networks take log-Mel spectrograms as input and consist of a sequence of convolutional layers with pooling applied along the time and frequency dimensions. A global pooling layer then summarizes the features into a fixed-length vector for classification.

For the downstream task, the resulting representation is passed to a task-specific classification layer. Instead of keeping the pre-trained backbone fixed, we fine-tune all model layers during training. This enables the learned representations to adapt to the target dataset while still benefiting from the information captured during pre-training.

We evaluate both CNN10 and CNN14 in order to examine the effect of model complexity on classification performance in our experiments.

6 Experimental Setup

6.1 Task definition

We consider a binary classification task between *asthma* and *non-asthma*. Each model takes a fixed-length audio segment, converts it into a log-Mel spectrogram, and outputs a prediction for one of the two classes.

Performance is evaluated at two levels: (i) **segment level**, where each audio segment is classified independently, and (ii) **patient level**, where predictions from multiple segments belonging to the same participant are aggregated to produce a single label per patient. We aggregate segment predictions by averaging the model outputs (logits) across all segments from the same participant and then taking the argmax to obtain a single patient label.

6.2 Models and fine-tuning strategy

Using the CNN10 and CNN14 variants, we assess pretrained convolutional audio models from the PANN family. For classification, a MLP head with two linear layers, dropout (Srivastava et al., 2014), and ReLU activation is added.

All models are fine-tuned end-to-end on the asthma classification task.

6.3 Optimization details

Training is implemented in PyTorch using Adam with weight decay. Using a single learning rate of 1×10^{-4} and weight decay of 1×10^{-4}, we

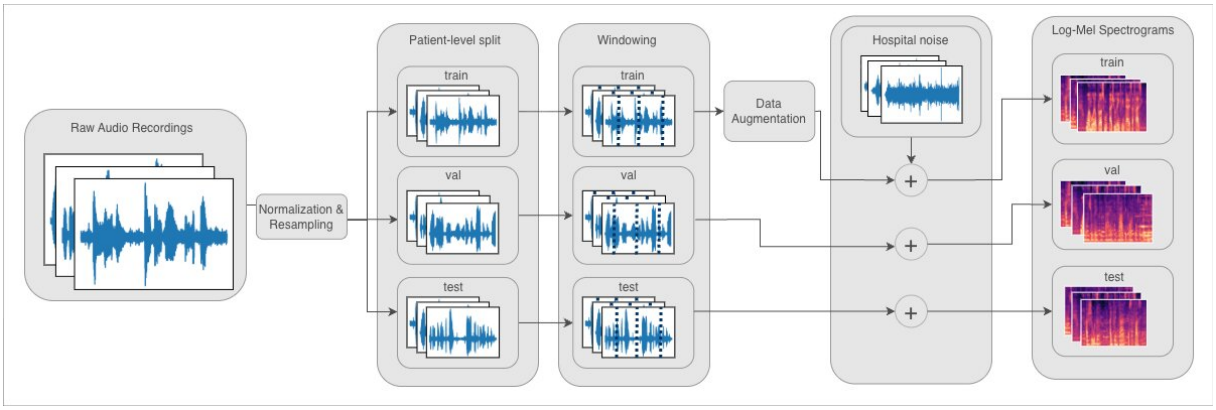


Figure 1: Overview of the audio preprocessing pipeline. Raw recordings are normalized and resampled, split at the participant level into training, validation, and test sets, segmented into fixed-length windows, augmented and mixed with environmental noise, and finally converted into time-frequency representations used as model inputs.

train for up to 50 epochs with batch size 16 and a fixed train/validation/test split of 60/20/20. Early stopping monitors validation, balancing accuracy with ten epochs of patience. Dropout ($p = 0.3$) is included in the MLP head.

6.4 Reproducibility

We fix random seeds and enable deterministic execution. All experiments are repeated with 10 seeds and reported as mean $\pm$ standard deviation.

7 Results

Table 3 reports accuracy and balanced accuracy for asthma vs. non-asthma classification at both segment and patient levels, reported as mean $\pm$ sample standard deviation across random seeds. Read speech yields higher performance than sustained vowels across models, with the strongest results obtained by CNN14 on speech.

Table 4 provides clinical metrics (sensitivity, specificity, and MCC) for all configurations, while Table 5 reports ROC-AUC and PR-AUC. Consistent with Table 3, speech-based models generally show stronger discrimination than vowel-based models.

Effect of pretraining (scratch baseline). We trained a scratch baseline with the same CNN10 Speech configuration but randomly initialized weights in order to separate the impact of extensive audio pretraining. The pretrained CNN10 Speech configuration outperforms the scratch model across Tables 3–5, suggesting that pretraining enhances generalization in this setting.

Patient-level ROC curves. Figure 2 visualizes patient-level ROC curves averaged across 10 random seeds. Pretrained models achieve higher ROC-AUC than the scratch baseline (Table 5), confirming the benefit of large-scale audio pretraining. CNN14 Speech achieves the highest ROC-AUC among pretrained models, closely followed by CNN10 Speech.

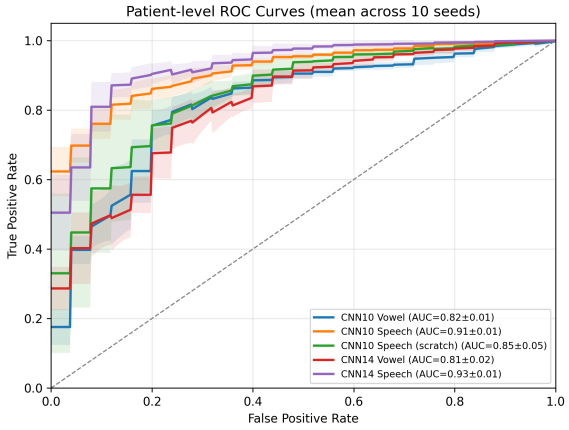


Figure 2: Patient-level ROC curves for the main speech-based configurations.

Statistical comparison (McNemar test). We use McNemar's exact test on paired *patient-level* predictions (Table 6). Speech vs. Vowel is significant for both CNN10 and CNN14 ($p < 0.001$) and pretrained vs. scratch, while CNN10 Speech vs. CNN14 Speech is not ($p = 0.454$).

Bootstrap confidence intervals (patient level). We estimate 95% bootstrap confidence intervals (1,000 patient resamples) for patient-level accuracy and balanced accuracy (Table 7). **CNN14 Speech** achieves the highest mean balanced accuracy, but its confidence interval overlaps with

Table 3: Sample mean ± standard deviation across random seeds for accuracy and balanced accuracy in asthma vs. non-asthma classification.

Model	Input	Segment		Patient	
		Acc	Bal	Acc	Bal
CNN10	Vowel	0.77 ± 0.03	0.77 ± 0.02	0.76 ± 0.03	0.76 ± 0.02
	Speech	**0.84 ± 0.02**	**0.80 ± 0.02**	**0.86 ± 0.01**	**0.81 ± 0.04**
	Speech (scratch)	0.73 ± 0.09	0.75 ± 0.04	0.73 ± 0.11	0.75 ± 0.05
CNN14	Vowel	0.73 ± 0.04	0.72 ± 0.03	0.73 ± 0.05	0.74 ± 0.03
	Speech	**0.84 ± 0.05**	**0.84 ± 0.02**	**0.85 ± 0.05**	**0.85 ± 0.03**

Table 4: Sample mean ± standard deviation across random seeds for sensitivity, specificity, and MCC in asthma vs. non-asthma classification.

Model	Input	Sens	Spec	MCC
CNN10	Vowel (Seg.)	0.77 ± 0.03	0.76 ± 0.03	0.49 ± 0.05
	Vowel (Pat.)	0.75 ± 0.04	0.78 ± 0.03	0.46 ± 0.03
	Speech (Seg.)	**0.86 ± 0.03**	**0.73 ± 0.07**	**0.55 ± 0.03**
	Speech (Pat.)	**0.90 ± 0.02**	**0.73 ± 0.09**	**0.61 ± 0.05**
	Speech (scratch, Seg.)	0.71 ± 0.14	0.78 ± 0.12	0.43 ± 0.07
	Speech (scratch, Pat.)	0.72 ± 0.17	0.77 ± 0.13	0.45 ± 0.08
CNN14	Vowel (Seg.)	0.74 ± 0.07	0.71 ± 0.07	0.41 ± 0.06
	Vowel (Pat.)	0.72 ± 0.07	0.75 ± 0.05	0.41 ± 0.06
	Speech (Seg.)	**0.84 ± 0.07**	**0.84 ± 0.06**	**0.60 ± 0.06**
	Speech (Pat.)	**0.85 ± 0.07**	**0.85 ± 0.07**	**0.65 ± 0.07**

Table 5: Sample mean ± standard deviation across random seeds for ROC-AUC and PR-AUC in asthma vs. non-asthma classification.

Model	Input	AUC	PR-AUC
CNN10	Vowel (Seg.)	0.82 ± 0.02	0.92 ± 0.01
	Vowel (Pat.)	0.82 ± 0.01	0.93 ± 0.00
	Speech (Seg.)	**0.91 ± 0.01**	**0.97 ± 0.00**
	Speech (Pat.)	**0.91 ± 0.01**	**0.98 ± 0.00**
	Speech (scratch, Seg.)	0.84 ± 0.04	0.95 ± 0.02
	Speech (scratch, Pat.)	0.85 ± 0.05	0.94 ± 0.03
CNN14	Vowel (Seg.)	0.80 ± 0.02	0.91 ± 0.01
	Vowel (Pat.)	0.81 ± 0.02	0.93 ± 0.01
	Speech (Seg.)	**0.92 ± 0.01**	**0.98 ± 0.00**
	Speech (Pat.)	**0.93 ± 0.01**	**0.98 ± 0.00**

Table 6: McNemar's exact test on patient-level predictions (two-sided), aggregated across 10 seeds using Fisher's method.

Comparison (A vs. B)	N	Discordant	p
CNN10 Speech vs. CNN10 Vowel	107	26.3	<0.001
CNN14 Speech vs. CNN14 Vowel	107	34.2	<0.001
CNN10 Speech (pretrained) vs. scratch	111	26.4	<0.001
CNN10 vs. CNN14 (Speech)	111	14.3	0.454
CNN10 vs. CNN14 (Vowel)	107	12.2	0.040

Table 7: Patient-level performance with 95% bootstrap confidence intervals (1,000 patient-resamples). Values are mean [95% CI] across 10 seeds.

Model	Acc	Bal. Acc
CNN10 Vowel	0.76 [0.68, 0.82]	0.76 [0.68, 0.84]
CNN10 Speech	0.86 [0.80, 0.91]	0.81 [0.73, 0.89]
CNN10 Speech (scratch)	0.74 [0.68, 0.79]	0.75 [0.68, 0.81]
CNN14 Vowel	0.73 [0.65, 0.79]	0.73 [0.65, 0.82]
CNN14 Speech	0.86 [0.80, 0.91]	0.86 [0.79, 0.91]

CNN10 Speech, indicating comparable performance across the two pretrained models.

7.1 Fairness Analysis across Demographic Groups

We evaluate potential demographic biases through a subgroup analysis based on age and sex using patient-level predictions, considering all model and input configurations. Because basic voice characteristics such as pitch and formant frequencies differ between males and females, this may influence the acoustic patterns learned by the models, motivating explicit evaluation across sex groups (Pépiot, 2015).

Participants were grouped by age into four bins (≤30, 31–45, 46–60, >60 years) and by sex. For each subgroup, we report classification accuracy.

Table 8 summarizes the results. Across both models, performance varies with age, with lower accuracy in the youngest groups and higher accuracy for participants aged 46 years and above. This

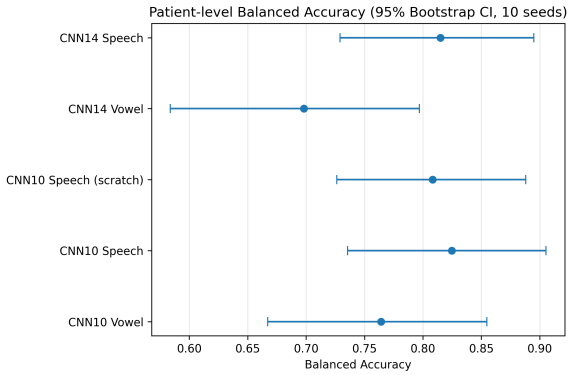


Figure 3: Patient-level balanced accuracy with 95% bootstrap confidence intervals for all evaluated models. CNN10 and CNN14 speech-based models achieve higher balanced accuracy than vowel-based models, while overlapping confidence intervals indicate comparable performance across several configurations.

effect is particularly pronounced for the Speech task, where both CNN10 and CNN14 achieve higher accuracy in older groups but perform poorly for younger participants. In contrast, the Vowel task exhibits more stable performance across age ranges, especially for CNN14, suggesting that sustained phonation may be less sensitive to age-related acoustic variability than read speech. However, estimates for the youngest age group should be interpreted with caution due to the small number of patients. Accuracy across sex is similar, although male participants are underrepresented, limiting statistical strength.

Overall, no strong bias is observed across sex, but the results indicate sensitivity to age differences and limited sample sizes.

8 Discussion

This work shows that pretrained convolutional audio models can effectively detect asthma from short voice and speech recordings. In all experiments, read speech outperformed sustained vowel phonation. This suggests that continuous speech contains richer information related to respiratory and phonatory behavior.

As mentioned in the Related Work section, previous studies have explored asthma detection from voice using handcrafted acoustic features and traditional machine learning models, mainly based on sustained vowels. Direct comparison with our study is not feasible due to differences in datasets, recording conditions, and languages. Nevertheless, our results are consistent with prior work in showing that asthma related information can be extracted from vocal signals. While sustained vowels were effective in earlier studies, our findings suggest that read speech provides even better discrimination. In contrast to these feature based approaches, we use pretrained convolutional audio models, highlighting the benefit of transfer learning.

Table 8: Sample mean ± standard deviation across random seeds for patient-level accuracy by age and sex groups.

Model	Input	Group	N	Acc
		Age groups		
CNN10	Vowel	≤30	4	0.38 ± 0.13
		31–45	9	0.84 ± 0.06
		46–60	15	0.71 ± 0.04
		>60	7	0.86 ± 0.12
	Speech	≤30	4	0.33 ± 0.12
		31–45	9	0.78 ± 0.13
		46–60	16	0.94 ± 0.05
		>60	8	0.99 ± 0.04
	Speech (scratch)	≤30	4	0.38 ± 0.24
		31–45	9	0.56 ± 0.23
		46–60	16	0.75 ± 0.17
		>60	8	0.80 ± 0.25
CNN14	Vowel	≤30	4	0.68 ± 0.24
		31–45	9	0.77 ± 0.10
		46–60	15	0.70 ± 0.13
		>60	7	0.76 ± 0.12
	Speech	≤30	4	0.38 ± 0.13
		31–45	9	0.67 ± 0.20
		46–60	16	0.89 ± 0.07
		>60	8	0.94 ± 0.11
		Sex		
CNN10	Vowel	Female	29	0.76 ± 0.03
		Male	6	0.75 ± 0.04
	Speech	Female	30	0.87 ± 0.01
		Male	7	0.81 ± 0.04
	Speech (scratch)	Female	91	0.74 ± 0.11
		Male	20	0.71 ± 0.11
CNN14	Vowel	Female	29	0.74 ± 0.05
		Male	6	0.69 ± 0.08
	Speech	Female	30	0.86 ± 0.05
		Male	7	0.83 ± 0.06

When comparing architectures, CNN10 and CNN14 exhibit comparable performance on read speech, with CNN14 achieving slightly higher discrimination metrics (e.g., ROC-AUC). This suggests that increasing model complexity does not necessarily yield consistent gains in accuracy in this setting. Given the current dataset size and overlapping confidence intervals, both architectures appear suitable, with CNN10 offering a simpler alternative and CNN14 benefiting from higher capacity in some metrics.

The comparison with a CNN10 trained from scratch highlights the importance of transfer learning. The pretrained CNN10 outperformed the ran-

domly initialized version, showing that large-scale audio pretraining is helpful when training data is limited. Even when the downstream task differs from the original objective, pretraining improves performance.

The fairness analysis showed differences across age groups, with lower accuracy for younger participants. This may be related to changes in speech with age and to data imbalance between groups. Performance across sex was more stable, although the small number of male participants limits stronger conclusions.

9 Conclusion

This work studied the use of pretrained neural audio models for asthma detection from voice and speech. By evaluating CNN10 and CNN14 models on vowel and read speech recordings, we showed that pretrained models perform better than models trained from scratch, highlighting the importance of transfer learning. The results also indicate that read speech provides stronger information for asthma detection than sustained vowels.

In addition, we analyzed performance at the patient level and across demographic groups. The results were stable across sex but varied across age groups, which emphasizes the importance of considering demographic factors in biomedical audio models.

Overall, the results confirm that pretrained audio models are a strong and effective approach for asthma detection from voice and speech. This work supports the potential of audio-based methods as a non-invasive tool for respiratory classification and motivates future studies with larger and more diverse datasets.

Limitations

This study has limitations. The dataset is relatively small and demographically imbalanced, which may limit generalization. We also consider only binary asthma classification, without severity or temporal modeling.

Ethical Considerations

This work uses voice and speech data from human participants collected with informed consent and handled in accordance with privacy and ethical guidelines. The work was approved by the Ethics Committee of Hospital (ommitted due to blind review). The data were anonymized, and the models are intended for research purposes only, not for clinical decision-making. We also examined demographic effects to help identify potential biases and support more responsible model development.

Acknowledgments

his study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior — Brasil (CAPES) — Finance Code 001. The work was conducted at the Center for Artificial Intelligence (C4AI-USP), with support from the University of São Paulo and the São Paulo Research Foundation (FAPESP) under grant #2019/07665-4.

Marcelo Finger was partly supported by the São Paulo Research Foundation (FAPESP) through grants 2023/00488-5 (SPIRA-BM) and 2022/11254-2 (EMU), and by the National Council for Scientific and Technological Development (CNPq) under grant PQ1 302963/2022-7.

The authors used generative AI tools to assist with writing (paraphrasing and language refinement) and code development. All AI-generated suggestions were verified and approved by the authors.

References

Kawther S Alqudaihi, Nida Aslam, Irfan Ullah Khan, Abdullah M Almuhaideb, Shikah J Alsunaidi, Nehad M Abdel Rahman Ibrahim, Fahd A Alhaidari, Fatema S Shaikh, Yasmine M Alsenbel, Dima M Alalharith, and 1 others. 2021. Cough sound detection and diagnosis using artificial intelligence techniques: challenges and opportunities. *Ieee Access*, 9:102327–102344.

E. Casanova, L. Gris, A. Camargo, D. da Silva, M. Gazzola, E. Sabino, A. Levin, A. Candido Jr, S. Aluisio, and M. Finger. 2021. Deep learning against covid-19: Respiratory insufficiency detection in brazilian portuguese speech. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021 (Online)*, pages 625–633. Association for Computational Linguistics.

Bradley E Chipps, Weily Soong, Reynold A Panettieri Jr, Warner Carr, Hitesh Gandhi, Wenjiong Zhou, Bill Cook, Jean-Pierre Llanos, and Christopher S Ambrose. 2023. Number of patient-reported asthma triggers predicts uncontrolled disease among specialist-treated patients with severe asthma. *Annals of Allergy, Asthma & Immunology*, 130(6):784–790.

Tusar Kanti Dash, Chinmay Chakraborty, Subhadip Mahapatra, and Ganapati Panda. 2022. Gradient boosting machine and efficient combination of features for speech-based detection of covid-19. *IEEE Journal*

of Biomedical and Health Informatics, 26(11):5364–5371.

Konstantinos P Exarchos, Maria Beltsiou, Chainti-Antonella Votti, and Konstantinos Kostikas. 2020. Artificial intelligence techniques in asthma: a systematic review and critical appraisal of the existing literature. *European Respiratory Journal*, 56(3).

Anne Fuhlbrigge, David Peden, Andrea J Apter, Homer A Boushey, Carlos A Camargo Jr, James Gern, Peter W Heymann, Fernando D Martinez, David Mauger, William G Teague, and 1 others. 2012. Asthma outcomes: exacerbations. *Journal of Allergy and Clinical Immunology*, 129(3):S34–S48.

Marcelo Matheus Gauy, Larissa Cristina Berti, Arnaldo Cândido, Augusto Camargo Neto, Alfredo Goldman, Anna Sara Shafferman Levin, Marcus Martins, Beatriz Raposo-de Medeiros, Marcelo Queiroz, Ester Cerdeira Sabino, Flaviane Romani Fernandes Svartman, and Marcelo Finger. 2023. Discriminant audio properties in deep learning based respiratory insufficiency detection in brazilian portuguese. In *Artificial Intelligence in Medicine: 21st International Conference on Artificial Intelligence in Medicine, AIME 2023, Portorož, Slovenia, June 12–15, 2023, Proceedings*, page 271–275, Berlin, Heidelberg. Springer-Verlag.

Marcelo Matheus Gauy and Marcelo Finger. 2022. Pretrained audio neural networks for speech emotion recognition in portuguese. In *First Workshop on Automatic Speech Recognition for Spontaneous and Prepared Speech Speech emotion recognition in Portuguese (SER 2022)*. CEUR-WS.

J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter. 2017. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 776–780. IEEE.

Murat Gezer, Mehmet Atilla Uysal, Neval Alagöz, Can Ortak, and Elif Yelda Niksarlıoğlu. 2025. Voices from the lungs: An innovative approach to asthma diagnosis using machine learning. *Acta Infologica*, 9(1):223–252.

Global Initiative for Asthma. 2024. Global strategy for asthma management and prevention. https://ginasthma.org.

M. A. Iqbal, Krishnamoorthy Devarajan, and Syed Musthak Ahmed. 2022. Real time detection and forecasting technique for asthma disease using speech signal and denn classifier. *Biomedical Signal Processing and Control*, 76:103637.

Md. Asim Iqbal, K. Devarajan, R. Lakshman Naik, and R. Sushmitha. 2024. Asthma detection from speech signals. In *Futuristic Trends in IoT*, volume 3 of *IIP Series*. IIP Series.

Panagiotis Kapetanidis, Fotios Kalioras, Constantinos Tsakonas, Pantelis Tzamalis, George Kontogiannis, Theodora Karamanidou, Thanos G Stavropoulos, and Sotiris Nikoletseas. 2024. Respiratory diseases diagnosis using audio analysis and artificial intelligence: a systematic review. *Sensors*, 24(4):1173.

Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley. 2020. Panns: Large-scale pretrained audio neural networks for audio pattern recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:2880–2894.

Maria Labied, Abdessamad Belangour, Mouad Banane, and Allae Erraissi. 2022. An overview of automatic speech recognition preprocessing techniques. In *2022 international conference on decision aid sciences and applications (DASA)*, pages 804–809. IEEE.

Sang Mee Lee, Hyein Ryu, Sunga Kong, Sun Hye Shin, Wooseong Huh, Myung Jin Chung, Juhee Cho, Taeyoung Kim, and Hye Yun Park. 2025. Voice as a digital biomarker: Foundation model-based copd assessment. *Research Square*. Preprint, under review at npj Digital Medicine.

Zhi Qing Looi, Zi Hao Ng, Rui Xiang Yak, Oren Rosen, and Anurag Kumar. 2024. Melodica ai-machine learning based detection of asthma via vocal audio analysis. In *Proceedings of the 2024 IEEE Conference on Artificial Intelligence (CAI)*, pages 905–910. IEEE.

Yi Lyu, Quan-Cheng Jiang, Shuai Yuan, Jing Hong, Chun-Feng Chen, Hai-Mei Wu, Yi-Qin Wang, Yu-Jing Shi, Hai-Xia Yan, and Jin Xu. 2025. Non-invasive acoustic classification of adult asthma using an xgboost model with vocal biomarkers. *Scientific Reports*, 15(1):28682.

Marcelo Matheus Gauy, Natália Hitomi Koza, Ricardo Mikio Morita, Gabriel Rocha Stanzione, Arnaldo Cândido Júnior, Larissa Cristina Berti, Anna Sara Shafferman Levin, Ester Cerdeira Sabino, Flaviane Romani Fernandes Svartman, and Marcelo Finger. 2026. Contrasting deep learning audio models for direct respiratory insufficiency detection versus blood oxygen saturation estimation. *Intelligence-Based Medicine*, 13:100331.

Daisuke Niizumi, Daiki Takeuchi, Masahiro Yasuda, Binh Thien Nguyen, Yasunori Ohishi, and Noboru Harada. 2025. Towards pre-training an effective respiratory audio foundation model. *arXiv preprint arXiv:2505.15307*.

Gaurav Pandey, Om P Pandey, Angela J Rogers, Mehmet E Ahsen, Gabriel E Hoffman, Benjamin A Raby, Scott T Weiss, Eric E Schadt, and Supinda Bunyavanich. 2018. A nasal brush-based classifier of asthma identified by machine learning analysis of nasal rna sequence data. *Scientific reports*, 8(1):8826.

Erwan Pépiot. 2015. Voice, speech and gender:. male-female acoustic differences and cross-language variation in english and french speakers. *Corela. Cognition, représentation, langage*, HS-16.

Yasir Rahmatallah, Aaron S. Kemp, Anu Iyer, Lakshmi Pillai, Linda J. Larson-Prior, Tuhin Virmani, and Fred Prior. 2025. Pre-trained convolutional neural networks identify parkinson's disease from spectrogram images of voice samples. *Scientific Reports*, 15(1):7337.

Om Prakash Singh, Ramaswamy Palaniappan, and MB Malarvili. 2018. Automatic quantitative analysis of human respired carbon dioxide waveform for asthma and non-asthma classification using support vector machine. *IEEE Access*, 6:55245–55256.

N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958.

Marko Topalovic, Stefan Laval, Jean-Marie Aerts, Thierry Troosters, Marc Decramer, Wim Janssens, Belgian Pulmonary Function Study investigators, and 1 others. 2017. Automated interpretation of pulmonary function tests in adults with respiratory complaints. *Respiration*, 93(3):170–178.

Laura Verde, Giuseppe De Pietro, and Giovanna Sannino. 2023. Artificial intelligence techniques for the non-invasive detection of covid-19 through the analysis of voice signals. *Arabian Journal for Science and Engineering*, 48(8):11143–11153.

Shengyun Wei, Shun Zou, Feifan Liao, and 1 others. 2020. A comparison on data augmentation methods based on deep learning for audio classification. *Journal of physics: Conference series*, 1453(1):012085.

Liang Xu, Lizhong Wang, Sijun Bi, Hanyue Liu, and Jing Wang. 2023. Semi-supervised sound event detection with pre-trained model. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Yun Xu and Royston Goodacre. 2018. On splitting training and validation set: a comparative study of cross-validation, bootstrap and systematic sampling for estimating the generalization performance of supervised learning. *Journal of analysis and testing*, 2(3):249–262.

Peisi Yin, Xiaoyu You, Xinyue Cui, Zhipeng Tang, Shanshan Yu, Huaian Fu, Fei Song, Kai Zhang, Xin Zhao, Lipeng Wang, and 1 others. 2025. Clinically diagnose asthma and monitor its severity using an ultrasensitive chemiresistive nitric oxide (no) gas sensor via exhaled breath analysis assisted by pattern recognition. *ACS sensors*.

Jun Zhan, Wen Chen, Longsheng Cheng, Qiong Wang, Feifei Han, and Yubao Cui. 2020. Diagnosis of asthma based on routine blood biomarkers using machine learning. *Computational intelligence and neuroscience*, 2020(1):8841002.

A RAG Chatbot with Incremental Context Retrieval based on Local LLMs for Hospital Documents

Murilo Vargas da Cunha[1,2] **Marília Rosa Silveira**[1] **César Brasil Sperb**[1]
Larissa Astrogildo Freitas[1] **Ulisses Brisolara Corrêa**[1]
[1]Federal University of Pelotas (UFPEL), Pelotas, RS, Brasil
[2]Federal Institute of Rio Grande do Sul (IFRS), Rio Grande, RS, Brasil
{mvcunha, mrsilveira, cbsperb, larissa, ulisses}@inf.ufpel.edu.br

Abstract

The adoption of LLMs in hospital environments demands solutions that ensure information security, computational efficiency, and rigorous control over sensitive institutional data. This work presents the development and evaluation of a chatbot based on RAG, using exclusively local LLMs, applied to internal documents of a university hospital in Portuguese, composed of Standard Operating Procedures and technical manuals. The methodology initially evaluates the quality of information retrieval through dense embedding models, measured by the Mean Reciprocal Rank (MRR) metric. Then, the generation stage is analyzed in two distinct scenarios: (i) RAG with fixed context, in which multiple chunks are provided simultaneously to the model, and (ii) Incremental page retrieval, in which chunks are sent sequentially according to the retrieval ranking. The generation assessment was conducted with four local LLMs — MedGemma3:27B, Gemma3:27B, Gpt-oss:20B, and Mistral Small 3.1 — using BERTScore as a quality metric. The results indicate that indiscriminate context increase in the fixed-context scenario degrades generation quality, even while increasing the probability of recovering the relevant chunk. In contrast, the incremental page retrieval technique showed improvements in BERTScore values, with the MedGemma3:27B model standing out with the best overall results. These findings demonstrate that adaptive context control is a critical factor in increasing the reliability and efficiency of RAG systems based on local LLMs in the healthcare domain.

1 Introduction

Large Language Models (LLMs) have been widely used in natural language processing (NLP) tasks in healthcare (Abo El-Enen et al., 2025; Lee et al., 2023). However, applications in the hospital setting involve highly sensitive documents, such as Standard Operating Procedures (SOPs), technical manuals, and internal regulations, which often cannot be sent to external services due to privacy restrictions, information security, and institutional compliance (Haltaufderheide and Ranisch, 2024; Li et al., 2023). In this scenario, the use of locally executed LLMs emerges as a viable alternative, as it allows data to remain entirely under the institution's control (Ng et al., 2024).

Nevertheless, these models are limited by their training data, which are often outdated and prone to generating inaccurate content, such as hallucinations, producing plausible but unfounded responses based on available data (Perković et al., 2024; Ji et al., 2023). This behavior represents a significant risk in hospital settings, where the accuracy and reliability of information are essential (Amugongo et al., 2025).

As an alternative to these implications, a technique has been widely adopted, Retrieval-Augmented Generation (RAG), in which the model generates responses conditioned on excerpts retrieved from an external document database (Neha et al., 2025; Oliveira et al., 2025). However, this strategy applied in local LLMs also presents additional limitations, especially related to the processing of large volumes of context, where the efficient incorporation of external knowledge during inference in long or dynamic contexts remains a challenge (Taguchi et al., 2025).

In this sense, fixed-context approaches tend to retrieve a predefined number of segments or chunks and concatenate them into a single input for the language model, which can introduce irrelevant or conflicting information. Furthermore, this strategy does not guarantee that the LLM will effectively utilize the retrieved data during generation, resulting in responses that may still be inaccurate or inconsistent (Asai et al., 2023).

This problem is particularly relevant in local LLMs, where a degradation in the quality of responses is observed as the size of the context in-

Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)
Vol. 2, pages 68–77

creases, whether due to attentional limitations, loss of focus, or a greater propensity for hallucinations (Levy et al., 2024; Li et al., 2024). Thus, the use of a fixed context may become inadequate in scenarios with extensive and heterogeneous documentary databases.

In order to mitigate this limitation, this work proposes an incremental page-by-page retrieval strategy, in which the retrieved chunks/pages are sent individually to the LLM, following the order of relevance ranking, interrupting the process as soon as a valid response is obtained. This approach avoids the cumulative sending of context and seeks to improve the effective use of retrieved information, reducing the interference of irrelevant segments. This method does not require fine-tuning of the model or additional inferences in LLM.

Given this context, this article presents the development and evaluation of a RAG-based chatbot using exclusively local LLMs, exploring the incremental page retrieval technique, applied to internal documents of a university hospital composed of SOPs and technical manuals in Portuguese. The evaluation considers both the quality of the retrieval and the generation of responses, contributing to the discussion of more effective and secure practices in the use of local LLMs in sensitive hospital environments.

2 Related Works

The use of RAG in healthcare applications has been extensively investigated as a strategy to mitigate hallucinations and increase the reliability of LLM-based systems. In hospital settings, the work of (Son et al., 2025) proposes a RAG chatbot to support operational queries in Electronic Patient Record (EMR) systems, employing contrast-learning-tuned multilingual embeddings and a commercial LLM via API. The results demonstrate high retrieval performance, with Top-K Accuracy exceeding 97%, focusing the analysis primarily on the retrieval stage.

In the context of patient education, the study by (Baur et al., 2025) presents a German-language RAG chatbot for orthopedics and traumatology, built from validated clinical guidelines and educational materials. The system combines semantic search with the Qdrant vector database and generation with OpenAI's GPT, being evaluated by automatic metrics and user studies, which indicate high acceptance and perceived quality. Another work that explores languages other than English is that of (Zhang et al., 2025a), in which the authors integrate RAG with medical knowledge graphs for clinical question-and-answer systems in Chinese, demonstrating consistent gains in accuracy in clinical benchmarks. These works reinforce the practical applicability of RAG in healthcare, without exploring adaptive mechanisms for context control.

Approaches focused on initial medical screening and guidance are also explored by (Nandi et al., 2024), who propose the MedMate chatbot, based on a RAG pipeline with BERT embeddings, retrieval via FAISS, and generation with LLaMA. The authors report an approximate accuracy of 76% and good alignment with medical recommendations, highlighting the potential of RAG to surpass traditional searches. The analysis, however, does not consider the impact of context size on the quality of the generation.

The work of (Kulshreshtha et al., 2025) develops a RAG-based medical chatbot using LLaMA 3.2-3B, LangChain, and FAISS, with a document-based database derived from MedlinePlus. The system injects retrieved snippets directly into the generator model's prompt and is evaluated using qualitative metrics of conciseness, accuracy, and relevance, as well as comparisons with commercial models. While demonstrating gains in reliability, the study does not analyze limitations associated with the use of extensive contexts.

Another study explores specialized biomedical models and semantic context enrichment; the authors of (Sinha et al., 2024) propose CMedRAGBot, which combines BioMistral-7B and PubMedBERT with conversational retrieval chains to support clinical triage. Although the study highlights the potential of RAG chatbots for clinical support and patient engagement, it does not systematically evaluate how the way context is delivered influences generation.

In contrast, the present work explicitly investigates the impact of context control on the quality of data generation in RAG systems applied to sensitive hospital documents, considering exclusively local LLMs.

3 Methodology

This section describes the methodology adopted for the development and evaluation of the chatbot. The methodology was structured in sequential steps, allowing for separate evaluation of information re-

trieval, response generation, and the proposed strategy for adaptive context control.

3.1 Dataset

The document collection consists of 107 internal normative documents from five hospital departments: Occupational Health and Safety Unit, Human Resources Division, Hospital Infection Control Service, Teaching and Research Management Department, and Information Technology and Digital Health Sector (Empresa Brasileira de Serviços Hospitalares (EBSERH), 2025).

The documents were segmented into smaller units (chunks) for indexing in the vector database. We adopted chunks of approximately 100 tokens, applying a soft limit: upon reaching this value, the cut was shifted to the next endpoint, preserving syntactic cohesion. To evaluate the generation of responses with LLM, the full text of the source page of each chunk was also stored in a dedicated field in the database containing approximately 400 tokens.

For the system evaluation, a hybrid dataset with 192 questions and answers was constructed, divided into two complementary subsets:

- **Synthetic dataset:** composed of 74 questions/answers automatically generated with Gemini 1.5 Flash, and manual identification of the identifier (ID) of the chunk containing the correct answer in the database. This subset allows for an objective and controlled evaluation of the retrieval step;
- **Expert dataset:** composed of 118 questions/answers developed by hospital professionals in a preliminary qualitative evaluation phase of the system.

Table 1 presents representative examples of questions and answers, illustrating the linguistic structure and informational granularity addressed in the evaluation.

3.2 Information Retrieval Assessment

The initial stage of the methodology evaluated the quality of information retrieval. This evaluation was performed using only the synthetic dataset of 74 questions, as it enables precise identification of whether the correct chunk was retrieved. The evaluated models were Gemma Embedding (embeddinggemma) (Vera et al., 2025), Qwen3 Embedding (qwen3embedding) (Zhang et al., 2025b), Granite Embedding (graniteembedding) (Awasthy et al., 2025), and multilingual-e5-small (intfloat/multilingual-e5-small) (Wang et al., 2024). Retrieval was performed using semantic similarity, measuring the cosine similarity between the questions and the chunks stored in the database, resulting in the selection of the 100 chunks with the best ranking. Performance was evaluated using the Mean Reciprocal Rank (MRR) metric. The evaluation process is shown in Figure 1.

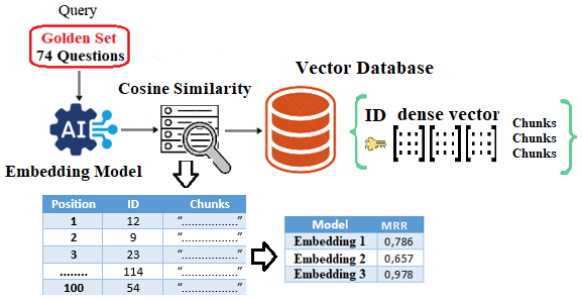


Figure 1: Information Retrieval Assessment Pipeline

3.3 Evaluation of Generation with Fixed Context

After defining the retrieval strategy, the generation stage was evaluated using the complete hybrid dataset of 192 questions, and the quality of the responses was assessed using the BERTScore metric, a semantic similarity metric based on contextual embeddings. The purpose of this stage was to evaluate how local LLMs behave when sent contexts containing a larger number of chunks in each inference. In this stage, the embedding model, qwen3embedding, was used, which showed better performance in the retrieval evaluation stage.

Four local LLMs were used, which were run on a machine equipped with an Nvidia RTX 5090 GPU: MedGemma3:27B, Gemma3:27B, Gpt-oss:20b, and Mistral Small 3.1. The impact of a fixed context size was analyzed using five independent scenarios, in which all questions in the dataset were submitted to the LLM using 1, 3, 5, 10, and 50 chunks/pages, respectively, corresponding to the top positions in the retrieval ranking. The objective of this step was to evaluate how increasing the context size affects the quality of the generation.

3.4 Incremental Page Retrieval Technique

As a main methodological contribution, this work presents a systematic analysis of the Incremental Page Retrieval technique, whose objective is to

Question	Answer
Qual o tempo médio de duração da higienização simples das mãos com água e sabonete, e qual o tempo médio da fricção antisséptica das mãos com preparações alcoólicas? (*What is the average duration of simple hand hygiene with soap and water, and what is the average time of antiseptic hand rubbing with alcohol-based preparations?*)	A higienização simples das mãos com água e sabonete deve durar em média 40-60 segundos; A fricção antisséptica das mãos dura em média 30 segundos. (*Simple hand hygiene with soap and water should last an average of 40-60 seconds; antiseptic hand rubbing lasts an average of 30 seconds.*)
Qual o tipo de acomodação para pacientes com tuberculose ativa? (*What type of accommodation is available for patients with active tuberculosis?*)	Pacientes com tuberculose ativa devem ser acomodados preferencialmente em quarto privativo, com portas fechadas e saída restrita. A exceção ocorre em casos de coorte, onde pacientes com a mesma patologia podem dividir o quarto, exceto se forem pacientes com suspeita ou confirmação de tuberculose resistente. (*Patients with active tuberculosis should preferably be accommodated in a private room with closed doors and restricted access. An exception is made in cohort cases, where patients with the same pathology may share a room, except if they are patients with suspected or confirmed drug-resistant tuberculosis.*)

Table 1: Example questions and answers

mitigate the quality degradation observed with the increase in cumulative context. In this approach, each chunk/page is sent individually to the LLM, avoiding the cumulative sending of context.

In the experiments, the language models were allowed to incrementally retrieve data up to the chunk/page at position 50, sending a single chunk/page at a time to the LLM, following the retrieval ranking order. The process is interrupted as soon as the model produces a response considered valid or when a maximum iteration limit is reached.

To enable this decision, the prompt explicitly instructs the LLM to indicate that the information was not found whenever the answer is not contained in the provided documents. Thus, an answer is classified as invalid when the model signals the absence of the information in the retrieved context. The identification of these answers was carried out using a mechanism based on regular expressions, designed to capture different linguistic variations used by the LLM to express negation or non-existence of the information. The response patterns were empirically defined from the observed outputs, allowing for the handling of textual variability and ensuring consistent detection of invalid answers.

Thus, experiments were conducted allowing incremental recovery up to the chunk at position 50, evaluating the efficiency of the method in terms of:

- Number of chunks needed to generate a valid response;
- Quality of the generated response (BERTScore).

This step was evaluated by combining the four LLM models (MedGemma3:27B, gemma3:27B, Mistral, and gpt-oss:20b) with the four embedding models (embeddinggemma, qwen3embedding, graniteembedding, and intfloat/multilingual-e5-small), totaling multiple experimental scenarios. The objective was to compare the retrieval efficiency between different LLM × embedding combinations using simple and interpretable metrics associated with the distribution of chunks used. The flowchart of this methodology is presented in Figure 2.

4 Results and Discussion

This section presents the results obtained in the experiments described in the methodology, followed by a comparative analysis and discussion of their implications.

4.1 Evaluation of Embedding Models in Information Retrieval

The results of the comparative analysis of the four embedding models are presented in Table 2, considering the synthetic dataset of 74 questions, in which the relevant chunk is previously known. The evaluation was conducted using the MRR metric, the median MRR, and the frequency of perfect retrieval (MRR = 1).

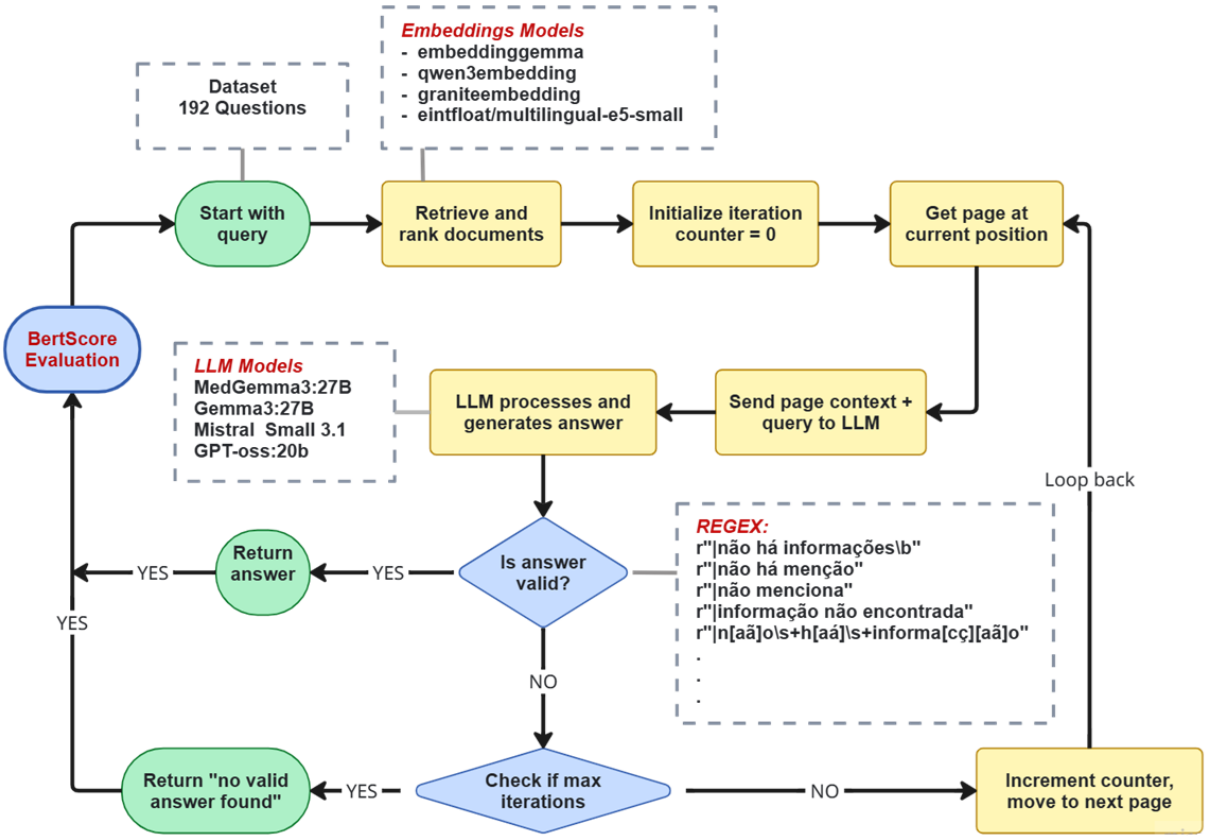


Figure 2: Flowchart of the proposed incremental retrieval pipeline

Model	Average	Median	Freq MRR = 1 (%)
qwen3embedding	0.8666	1	79.73
embeddinggemma	0.8348	1	72.97
graniteembedding	0.7695	1	66.22
intfloat/multilingual	0.6782	1	52.70

Table 2: MRR metric statistics by model

Among the models tested, qwen3embedding showed the best overall performance, with an average MRR of 0.8666 and a perfect recovery rate of 79.73%, indicating a greater ability to position the relevant chunk at the top of the ranking. In comparison, embeddinggemma showed intermediate performance, with an average MRR of 0.8348 and perfect recovery in 72.97% of queries. Graniteembedding and multilingual-e5-small obtained the lowest values, with average MRRs of 0.7695 and 0.6782, and perfect recovery rates of 66.22% and 57.70%.

Overall, the results indicate that the qwen3embedding and embeddinggemma models performed better, as the relevant chunk is retrieved in the first position in most queries.

4.2 Results of the Generation Evaluation with Fixed Context

This stage presents the results of the generation evaluation using a fixed context, in which a predefined number of retrieved chunks/pages, ranked by cosine similarity, is sent to the language model, regardless of the actual information needed to answer the question. The experiments were conducted in five scenarios, varying the number of chunks/pages sent to the model (1, 3, 5, 10, and the first 50 chunks/pages from the retrieval ranking). The quality of the responses was evaluated using the BERTScore (F1) metric.

The results are presented in Table 3 and show a consistent behavior among all the models evaluated: the quality of the generation progressively decreases as the context size increases.

Model	1 Page	3 Pages	5 Pages	10 Pages	50 Pages
MedGemma3:27B	0.8058	0.7991	0.7694	0.6841	0.6575
Gemma3:27B	0.8020	0.7958	0.7704	0.6956	0.6634
Gpt-oss:20b	0.7674	0.7601	0.7539	0.7098	0.6401
Mistral Small 3.1	0.7947	0.7820	0.7388	0.6726	0.6541

Table 3: BertScore(F1) results fixed context

The degradation becomes more pronounced start-

ing at 10 pages, and even more evident in the extreme scenario of 50 pages, where all models show a substantial drop in performance. In the case of gemma3:27B, for example, increasing the context from 1 to 50 chunks results in an approximate 17% reduction in the BERTScore. A similar trend is observed for the other models, indicating that this behavior is not specific to a particular architecture.

These results suggest the occurrence of a context dilution effect, in which the excessive inclusion of irrelevant information or information only partially related to the question hinders the model's ability to identify the truly useful sections for generating the answer. Although increasing the number of chunks/pages raises the probability that the correct section is present in the context, this benefit is outweighed by the negative impact of information overload.

Thus, the results of this stage provide a clear empirical justification for the proposed adaptive context control strategy presented in this work, which aims to balance informational coverage and generation quality, avoiding the unnecessary sending of large volumes of context to the language model.

4.3 Results of Incremental Page Retrieval

This subchapter presents and discusses the results of the Incremental Page Retrieval strategy, proposed in this work as an alternative to fixed context delivery. To better summarize and understand the results obtained, Table 4 is presented, comparing different combinations of LLMs and embeddings, showing performance in generation and retrieval quality metrics. The BERTScore (F1), percentages of responses found with 1, up to 3, and up to 5 chunks, as well as the maximum retrieval ranking required to obtain a valid response, are presented. Finally, the percentage of cases not found indicates the coverage and efficiency of each configuration.

The results show that the vast majority of questions were answered using only the first chunk/page retrieved, regardless of the language model or embedding employed. In the case of Gpt-oss:20b, the proportion of questions answered with only one chunk ranged from 79.69% to 91.15%, with qwen3embedding showing the highest percentage. Similar behavior was observed for gemma3:27B, whose values ranged from 77.08% to 90.10%, again with better performance associated with qwen3embedding. Mistral small 3.1 showed slightly lower, but still high, percentages, ranging from 70.31% to 84.38%.

However, there are cases that required the system to perform a deeper search in the ranking, for example, the pair (Mistral S3.1 + intfloat/multilingual) that went up to the chunk in position 50 to answer a question, suggesting that the incremental retrieval system is robust enough to recover information that the initial ranking places in unfavorable positions, which would be lost in a RAG system with a fixed and low k (e.g., $k = 5$).

Another factor observed is that even when using the same embedding model and, therefore, the same retrieval ranking, the different LLM models arrive at the answer in distinct maximum chunk positions. This behavior indicates that, for certain questions, the models consider different contexts as sufficient throughout the ranking, despite the ordering of the retrieved documents being identical.

Table 4 also reports the average BERTScore (F1). The incremental retrieval strategy consistently maintains or improves response quality compared to fixed-context configurations using multiple chunks. Medgemma3:27B, for example, presented the best overall results, reaching 0.8124 with embeddinggemma and 0.8116 with qwen3embedding. As for most queries, the models considered that the first chunk contained sufficient information to generate a response, this reinforces the hypothesis that sending large volumes of context is unnecessary in most cases where there is a good information retrieval step.

Nevertheless, it is important to emphasize that a central aspect of the proposed methodology is the joint evaluation of the quality of generation and the behavior of incremental retrieval. By simultaneously analyzing the BERTScore of the generated responses and the distribution of the number of chunks needed to answer each question, it becomes possible to identify scenarios where the model prematurely interrupts retrieval, incorrectly assuming that the relevant information has been found. In these cases, although the chunk distribution indicates high efficiency, often with responses generated from the first chunk, the BERTScore shows semantic degradation in relation to the answer key, characterizing inaccurate or misleading responses.

This behavior can be observed in the Gpt-oss:20b model, which presents a favorable distribution of the number of chunks used, with a high frequency of responses generated from the first chunks in the ranking. However, this behavior is accompanied by lower BERTScore values than the other models evaluated.

Table 4: Average performance results of LLMs considering different embedding models

Model (LLM + Embedding)	BertScore(F1)	% 1 Chunk	% ≤ 3 Chunks	% ≤ 5 Chunks	Chunk Máx	% Not Found
MedGemma 3 27B + Emb Gemma	0.8124	87.50%	94.79%	95.31%	19º Position	1.04%
MedGemma 3 27B + Qwen 3 8B	0.8116	87.50%	96.35%	97.39%	21º Position	1.04%
MedGemma 3 27B + Granite 278M	0.8009	83.33%	94.26%	95.30%	26º Position	0.00%
MedGemma 3 27B + intfloat/multilingual	0.7916	76.56%	89.59%	94.79%	21º Position	0.52%
Gemma 3 27B + Emb Gemma	0.8087	88.54%	95.31%	96.35%	19º Position	0.00%
Gemma 3 27B + Qwen 3 8B	0.8090	90.10%	94.79%	96.87%	47º Position	0.00%
Gemma 3 27B + Granite 278M	0.7942	86.46%	94.27%	95.83%	26º Position	0.52%
Gemma 3 27B + intfloat/multilingual	0.7924	77.08%	87.50%	92.19%	44º Position	0.52%
GPT OSS 20B + Emb Gemma	0.7678	87.50%	95.83%	96.35%	16º Position	0.00%
GPT OSS 20B + Qwen 3 8B	0.7635	91.15%	96.88%	98.96%	7º Position	0.00%
GPT OSS 20B + Granite 278M	0.7489	86.46%	95.31%	96.88%	17º Position	0.00%
GPT OSS 20B + intfloat/multilingual	0.7520	79.69%	93.24%	97.40%	38º Position	0.00%
Mistral S3.1 + Emb Gemma	0.7988	82.81%	94.27%	95.31%	16º Position	0.00%
Mistral S3.1 + Qwen 3 8B	0.7970	84.38%	93.75%	96.35%	35º Position	0.00%
Mistral S3.1 + Granite 278M	0.7880	79.69%	91.67%	92.71%	26º Position	0.52%
Mistral S3.1 + intfloat/multilingual	0.7887	70.31%	85.94%	90.62%	50º Position	0.52%

To corroborate this observation, a two-dimensional analysis is presented in Figure 3, which reveals that efficiency in incremental retrieval is not linearly linked to semantic quality. While the Gemma:3 27B model demonstrates superiority on both axes, it is observed that gpt-oss:20b, despite reaching the stopping criterion (valid response) early in 91.15% of cases, presents an average BERTScore 5.6% lower than MedGemma:3 27B under the same embedding conditions. This behavior indicates that the validation of a 'valid response' in the incremental process can be achieved by simpler models with less precise content, reinforcing the importance of evaluating the generated responses with metrics such as the BertScore for quality control in RAG systems with dynamic retrieval.

Finally, the analysis of the results reinforces the limitations of traditional approaches based on fixed top-K and positions Incremental Page Retrieval as an effective adaptive strategy, particularly suitable for local LLM scenarios where the retrieval step is not as efficient and the chunk with the context of the correct answer is located further down in the ranking.

5 Conclusions and Future Work

This work presented the development and evaluation of a RAG-based chatbot using exclusively local LLMs, applied to institutional documents of a university hospital in Portuguese. The results showed that, although information retrieval has high performance, the indiscriminate increase in context sent to the model compromises the quality of the generation, as indicated by the consistent reduction in the BERTScore. This behavior reinforces the need for strategies that dynamically control the use of context in RAG systems, especially in sensitive domains such as healthcare.

As a main contribution, the incremental page retrieval technique was proposed and evaluated, in which chunks are sent individually to the LLM following the retrieval ranking, interrupting the process as soon as a valid answer is obtained. The results demonstrate that most questions can be answered with only the first or second chunk/page, significantly reducing the volume of context, the number of model calls, and the computational cost, without compromising the semantic quality of the answers. The joint analysis of the BERTScore and the chunk distribution proved fundamental in identifying hallucination behaviors and qualitative differences between the evaluated models.

Another noteworthy factor was the improvement in BertScore(F1) results after applying the incremental page retrieval technique, with particular emphasis on the performance of the MedGemma:3 27B model in generating responses, which proved superior to the other models. This can be attributed to the fact that its training was performed with text data and medical images. In addition to the empirical gain in response quality, the proposed strategy introduces flexibility to retrieve subsequent chunks when necessary, allowing the system to retrieve relevant information positioned deeper in the ranking, without incurring the cumulative context expansion observed in fixed-top-K approaches.

Although the results obtained are indicative of the behavior of the evaluated models, some limitations should be considered. Firstly, the retrieval

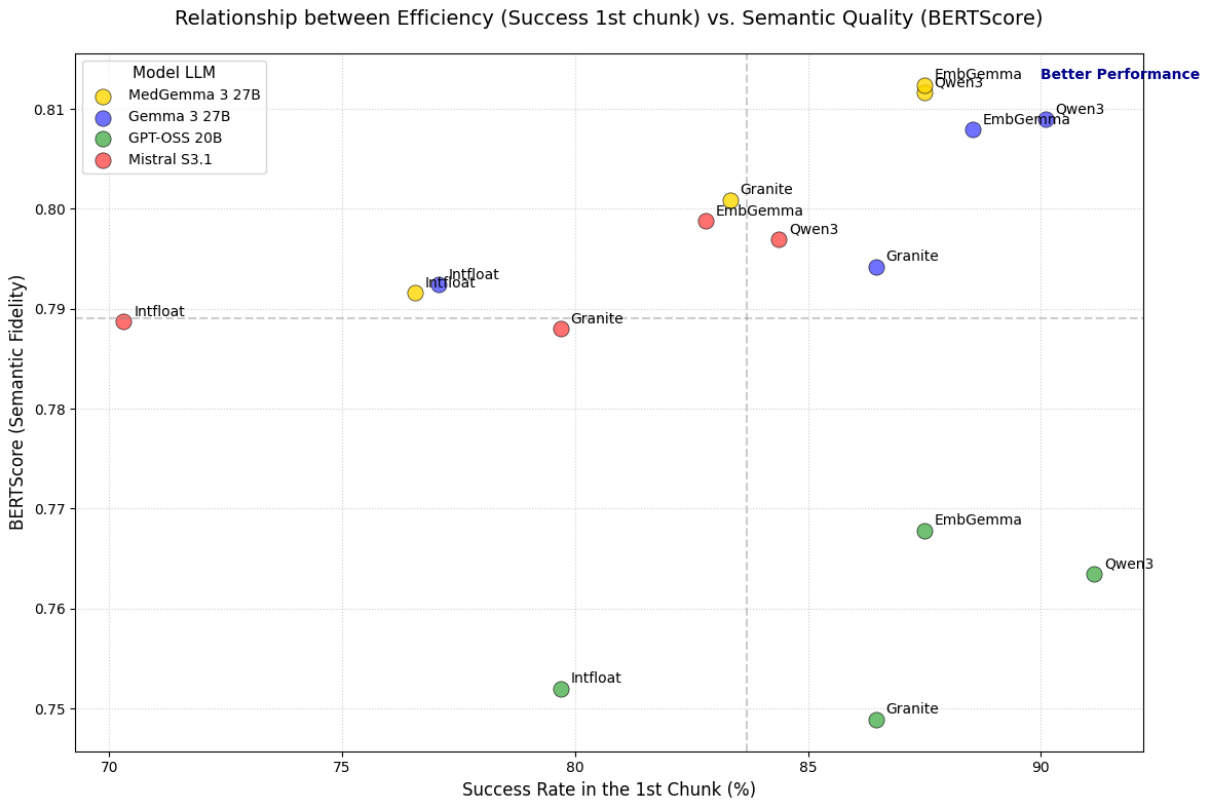


Figure 3: Scatter plot correlating First-Chunk Success and BERTScore for each pair.

assessment is based on a synthetic dataset composed of 74 questions, which, despite allowing for the precise identification of the relevant chunk, may not fully reflect the linguistic and semantic diversity observed in real-world usage scenarios. Similarly, context progression strategies, in which multiple chunks are incrementally accumulated to compose a more complete response, were not evaluated. Therefore, questions whose answer explicitly depends on the combination of information distributed across more than one chunk may not be fully covered by the adopted method.

Future work will involve investigating the progressive context retrieval technique, in which the number of chunks is expanded only when the model explicitly indicates the absence of relevant information, seeking to mitigate cases where the response depends on multiple complementary chunks. Additionally, a qualitative evaluation with hospital end-users, including healthcare professionals, is planned to analyze aspects such as usefulness, clarity, reliability, and adequacy of responses in real-world use. Finally, future extensions include the analysis of other LLM models and embeddings, as well as the integration of automatic mechanisms for detecting uncertainty and hallucination.

Limitations

Although hybrid retrieval approaches such as BM25-based methods, rerankers, and commercial LLM baselines have shown strong performance in RAG systems, these configurations were not explored in the present study due to the scope and space limitations of this work. The primary focus of this paper is the evaluation of RAG pipelines using locally deployed LLMs in healthcare-related datasets, motivated by privacy and data protection constraints.

Experiments involving hybrid retrieval strategies, BM25-based ranking, rerankers, and comparisons with commercial LLMs were previously investigated in our earlier study (da Cunha et al., 2025), where a broader benchmarking of RAG configurations was conducted. Therefore, the present work complements that study by focusing specifically on privacy-preserving RAG architectures using local LLMs in sensitive data environments.

Acknowledgments

This work was supported by Instituto Federal do Rio Grande do Sul (IFRS), Empresa Brasileira e Serviços Hospitalares (EBSERH) and Hospital Es-

cola da UFPEL. This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001. We would like to thank the FAPERGS - Brasil for Financial Support, Award Agreement 22/2551-0000598-5. We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan X Pascal GPU used for this research.

References

Mohamed Abo El-Enen, Sally Saad, and Taymoor Nazmy. 2025. A survey on retrieval-augmentation generation (rag) models for healthcare applications. *Neural Computing and Applications*, 37(33):28191–28267.

Lameck Mbangula Amugongo, Pietro Mascheroni, Steven Brooks, Stefan Doering, and Jan Seidel. 2025. Retrieval augmented generation for large language models in healthcare: A systematic review. *PLoS Digital Health*, 4(6).

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2023. Self-RAG: Self-reflective retrieval augmented generation. In *NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following*.

Parul Awasthy, Aashka Trivedi, Yulong Li, Mihaela Bornea, David Cox, Abraham Daniels, Martin Franz, Gabe Goodhart, Bhavani Iyer, Vishwajeet Kumar, Luis Lastras, Scott McCarley, Rudra Murthy, Vignesh P, Sara Rosenthal, Salim Roukos, Jaydeep Sen, Sukriti Sharma, Avirup Sil, and 3 others. 2025. Granite embedding models. *Preprint*, arXiv:2502.20204.

David Baur, Jörg Ansorg, Christoph-Eckhard Heyde, and Anna Voelker. 2025. Development and evaluation of a retrieval-augmented generation chatbot for orthopedic and trauma surgery patient education: Mixed-methods study. *JMIR AI*, 4:e75262.

Murilo Vargas da Cunha, Marilia Rosa Silveira, Brenda Salenave Santana, Larissa Astrogildo Freitas, and Ulisses Brisolara Corrêa. 2025. Optimizing and evaluating a retrieval-augmented generation system for normative document retrieval in hospital settings. In *Brazilian Symposium on Multimedia and the Web (WebMedia)*, pages 385–393. SBC.

Empresa Brasileira de Serviços Hospitalares (EBSERH). 2025. Estrutura administrativa — hu-ufsc. https://www.gov.br/ebserh/pt-br/hospitais-universitarios/regiao-sul/hu-ufsc/governanca/estrutura-administrativa. Acesso em: 13 jul. 2025.

Joschka Haltaufderheide and Robert Ranisch. 2024. The ethics of chatgpt in medicine and healthcare: a systematic review on large language models (llms). *NPJ digital medicine*, 7(1):183.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12):38.

Agrim Kulshreshtha, Aditya Choudhary, Tejas Taneja, and Seema Verma. 2025. Enhancing healthcare accessibility: A rag-based medical chatbot using transformer models. In *2024 International Conference on IT Innovation and Knowledge Discovery (ITIKD)*, pages 1–4. IEEE.

Peter Lee, Carey Goldberg, and Isaac Kohane. 2023. *The AI revolution in medicine: GPT-4 and beyond*. Pearson.

Mosh Levy, Alon Jacoby, and Yoav Goldberg. 2024. Same task, more tokens: the impact of input length on the reasoning performance of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15339–15353.

Hanzhou Li, John T Moon, Saptarshi Purkayastha, Leo Anthony Celi, Hari Trivedi, and Judy W Gichoya. 2023. Ethics of large language models in medicine and medical research. *The Lancet Digital Health*, 5(6):e333–e335.

Jiaqi Li, Mengmeng Wang, Zilong Zheng, and Muhan Zhang. 2024. Loogle: Can long-context language models understand long contexts? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16304–16333.

Avhishek Nandi, Barnali Paul, Piyali Datta, and Deepsubhra Guha Roy. 2024. Medmate: A contextual approach for disease diagnosis using retrieval-augmented generation. In *International Conference on Recent Advances in Artificial Intelligence & Smart Applications*, pages 429–441. Springer.

Fnu Neha, Deepshikha Bhati, and Deepak Kumar Shukla. 2025. Retrieval-augmented generation (rag) in healthcare: A comprehensive review. *AI*, 6(9):226.

Karen Ka Yan Ng, Izuki Matsuba, and Peter Chengming Zhang. 2024. Rag in health care: A novel framework for improving communication and decision-making by addressing llm limitations. *NEJM AI*.

S. S. T. Oliveira, D. Fazzioni, and D. O. C. Ferreira. 2025. Grandes modelos de linguagem. In *In: Kudo, T. N. et al.* Cegraf UFG, Goiânia. E-book (254 p.). ISBN 978-85-495-1096-9.

G. Perković, A. Drobnjak, and I. Botički. 2024. Hallucinations in llms: Understanding and addressing challenges. In *Proceedings of the 47th MIPRO ICT and Electronics Convention (MIPRO 2024)*, pages 2084–2088, Opatija, Croatia. IEEE.

Kushagra Sinha, Vaibhav Singh, Ankit Vishnoi, Parul Madan, and Yadvendra Shukla. 2024. Healthcare diagnostic rag-based chatbot triage enabled by biomistral-7b. In *2024 International Conference on Emerging Technologies and Innovation for Sustainability (EmergIN)*, pages 333–338. IEEE.

Namrye Son, Inchul Kang, Inhu Kim, Keehyuck Lee, Sejin Nam, and Donghyoung Lee. 2025. Development and evaluation of a retrieval-augmented generation-based electronic medical record chatbot system. *Healthcare Informatics Research*, 31(3):218–225.

Chihiro Taguchi, Seiji Maekawa, and Nikita Bhutani. 2025. Efficient context selection for long-context qa: No tuning, no iteration, just adaptive-k. *arXiv preprint arXiv:2506.08479*.

Henrique Schechter Vera, Sahil Dua, Biao Zhang, Daniel Salz, Ryan Mullins, Sindhu Raghuram Panyam, Sara Smoot, Iftekhar Naim, Joe Zou, Feiyang Chen, Daniel Cer, Alice Lisak, Min Choi, Lucas Gonzalez, Omar Sanseviero, Glenn Cameron, Ian Ballantyne, Kat Black, Kaifeng Chen, and 70 others. 2025. Embeddinggemma: Powerful and lightweight text representations. *Preprint*, arXiv:2509.20354.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Multilingual e5 text embeddings: A technical report. *arXiv preprint arXiv:2402.05672*.

Dongfang Zhang, Haoze Du, Xiaolei Wang, Mingdong Zhu, Xiaoxiao Pang, Dongqing Wei, and Xianfang Wang. 2025a. Cmedragbot: A chinese medical chatbot based on graph rag and large language models. *Interdisciplinary Sciences: Computational Life Sciences*, pages 1–16.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025b. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*.

A Dataset of Brazilian Portuguese Clinical Notes for Anaphylaxis Detection

Matheus Machado and Vinícius Vanzin and Dilvan Moreira
Institute of Mathematics and Computer Science
University of São Paulo
matheusmatos@usp.br, vinicius.vanzin@usp.br, dilvan@icmc.usp.br

Luis Felipe Ensina and Fábio Lario
Dep. of Allergy and Clinical Immunology
Hospital Sírio-Libanês
100alergia@gmail.com, fabio.clario@hsl.org.br

Abstract

Anaphylaxis is an acute, potentially life-threatening allergic reaction that requires rapid recognition in clinical settings. Natural language processing (NLP) approaches for automatic detection of anaphylaxis in clinical narratives can support large-scale analysis of health records and retrospective clinical research. However, such approaches depend on high-quality labeled corpora, and resources for Portuguese remain scarce. This paper introduces a corpus of Brazilian Portuguese clinical notes annotated by domain specialists for the presence or absence of anaphylaxis. The dataset comprises 969 clinical narratives drawn from three sources: clinician-authored synthetic clinical notes designed to represent realistic scenarios, case reports from the medical literature rewritten into note-like format by specialists, and a subset of de-identified notes from the publicly available SemClinBr corpus. All texts were reviewed and labeled by allergists using established clinical diagnostic criteria, and the corpus reflects realistic prevalence conditions, with approximately 5% positive cases. We describe the corpus design, data sources, annotation methodology, and composition, discuss potential research applications, and address ethical considerations. The corpus is intended as a reusable resource for Portuguese clinical NLP, supporting future work on document classification, information extraction, and language modeling in the medical domain.

1 Introduction

Anaphylaxis is an acute, potentially life-threatening systemic hypersensitivity reaction that may progress rapidly without prompt treatment. In clinical practice, its recognition relies on patient history, physical examination, and judgment guided by standardized diagnostic criteria. The National Institute of Allergy and Infectious Disease/Food Allergy and Anaphylaxis Network (NIAID/FAAN) criteria define anaphylaxis through symptom patterns involving acute mucocutaneous manifestations with respiratory or cardiovascular compromise, rapid multi-system involvement after exposure to a likely allergen, or hypotension after exposure to a known allergen (Dribin et al., 2023). The World Allergy Organization (WAO) 2020 update simplifies these criteria and highlights additional presentations, including severe gastrointestinal symptoms and cases dominated by respiratory or cardiovascular signs (Cardona et al., 2020). Despite these guidelines, diagnosis remains challenging because clinical presentations vary across patients and episodes, and documentation may be incomplete or heterogeneous in real-world records (Simons et al., 2011).

Automatic detection of anaphylaxis in clinical narratives could support research on case identification at scale and facilitate retrospective analyses of electronic health records. However, supervised approaches depend on annotated corpora, and clinical text resources remain limited in languages other than English. In Brazilian Portuguese, SemClinBr is one of the few publicly available clinical corpora; it comprises 1,000 de-identified notes annotated with 65,117 entities and 11,263 relations (Oliveira et al., 2022). While SemClinBr is valuable for tasks such as named-entity recognition and relation extraction, it does not provide document-level labels for specific conditions such as anaphylaxis. More broadly, access to clinical narratives is constrained by privacy and governance requirements, motivating increased attention to dataset transparency

Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)
Vol. 2, pages 78–87

through documentation practices such as dataset statements and datasheets (Bender and Friedman, 2018; Gebru et al., 2021). Synthetic clinical text has also been explored as a complementary strategy to mitigate data scarcity and reduce privacy risks by generating plausible narratives that do not correspond to identifiable individuals (Mendes et al., 2025).

In this work, we introduce a dataset of Brazilian Portuguese clinical notes annotated for the presence or absence of anaphylaxis. The contribution is dataset-centric: we focus on documenting the corpus design, data sources, annotation criteria, and corpus characteristics to support reproducible research on Portuguese clinical NLP. The dataset combines expert-authored synthetic clinical notes, adapted case reports rewritten into a note-like format, and de-identified notes sampled from a public clinical corpus. We further discuss limitations and ethical considerations relevant to releasing and using a clinical text resource. The corpus is available publicly at Hugging Face: matos1012/brazilian-portuguese-anaphylaxis.

2 Related Work

Research on automatic processing of clinical text encompasses the development of linguistic resources, the application of NLP methods to clinically relevant phenomena, and approaches to mitigate data scarcity and privacy constraints. While these areas are well established for English-language clinical NLP, corresponding resources and studies remain limited for Portuguese. In this context, the availability of clearly documented datasets is essential for enabling reproducible and comparable research. This section reviews work most relevant to the present study, focusing on clinical corpora in Portuguese, NLP-based approaches to anaphylaxis detection, and the use of synthetic data in clinical NLP.

2.1 Clinical corpora in Portuguese

Publicly available resources for Portuguese clinical NLP are historically scarce, a disparity that becomes evident when compared to mature English-language benchmarks. Examples of large-scale datasets include the MIMIC-IV-Note collection, which aggregates over 331,000 de-identified discharge summaries from the Beth Israel Deaconess Medical Center (Johnson et al., 2023), and the gold-standard annotated corpora from the n2c2 (formerly i2b2) shared tasks, typically comprising 500 to 1,300 specialized notes per task (Wang et al., 2020).

In the Portuguese clinical NLP landscape, early efforts focused on specialized domains, such as the publicly-available CLINpt corpus (Lopes et al., 2019), which provided 281 neurology case reports from medical journals. Schneider et al. (2020) later utilized a private collection of 3.8 million sentences derived from Brazilian hospital EHRs and scientific abstracts to develop the BioBERTpt model. Significant progress was made with SemClinBr (Oliveira et al., 2022), the first large-scale multi-institutional corpus for Brazilian Portuguese, offering 1,000 notes semantically annotated with over 65,000 entities. Concurrently, the credentialed-access BRATECA dataset (Dias and Ulbrich, 2022) provided a leap in volume, aggregating over 2.5 million free-text notes from ten Brazilian hospitals, while da Rocha et al. (2023) introduced the private HCFMB corpus, consisting of 1,200 clinical texts derived from 30,000 records at a university hospital.

More recent contributions include AnonyMED-BR (Schiezaro et al., 2025), a publicly-available dataset of 2,962 records specifically designed for anonymization tasks, and the anonymized sepsis summaries from da Silva and Pazin-Filho (2025), featuring 200 long-form discharge summaries from a tertiary hospital. MedPT (Färber et al., 2025) represents a consumer health-oriented dataset, containing over 384,000 authentic question-answer pairs from patient-doctor interactions. Clinical NLP research in Portuguese frequently relies on private corpora, limited-access resources, or translated datasets, underscoring the need for additional publicly documented datasets targeting clinically meaningful phenomena.

2.2 Anaphylaxis detection using NLP

Anaphylaxis detection in clinical text is a complex phenotyping task that has evolved from rule-based systems to the use of large language models (LLMs). Early work by Botsis et al. (2012) utilized the VAERS dataset to extract features from thousands of vaccine safety reports using semantic text mining. Botsis and Ball (2013) furthered this by automating case definitions through literature-based reasoning, relying on mapping synonyms to Brighton criteria. Walsh et al. (2013) highlighted the limitations of structured data using a critical benchmark of 122 potential anaphylaxis events

across eight healthcare settings. Segura-Bedmar et al. (2018) shifted toward large-scale detection, applying convolutional neural networks and classical classifiers to a collection of over 219,000 clinical records.

Recent studies have emphasized the integration of clinical notes with structured records or the application of instruction-tuned models. Yu et al. (2020) developed algorithms for the Vaccine Safety Datalink, utilizing 311 potential cases to identify vaccine-related anaphylaxis with high specificity. Lo et al. (2022) addressed allergy reconciliation by detecting discrepancies in encounter notes from the Mass General Brigham healthcare system. Carrell et al. (2023) engineered a set of NLP-derived symptom and criteria covariates using data from 516 patients across the Kaiser Permanente network. In the Portuguese context, Machado et al. (2024) assessed several large language models on a smaller set of annotated clinical reports, focusing on model behavior under limited data conditions. Finally, Ensina et al. (2025) evaluated LLMs on a corpus of Portuguese medical texts annotated by physicians for anaphylaxis, exploring different prompting strategies and reporting high classification performance.

The present work differs in scope by focusing on the dataset itself. Rather than reporting additional model results, we document the construction, annotation methodology, and characteristics of the corpus to support reuse and transparent evaluation.

2.3 Synthetic clinical data

Synthetic clinical data generation has emerged as a key strategy to circumvent privacy constraints and data scarcity. Early methods, such as SynthNotes (Begoli et al., 2018), utilized semantic templates to generate note-like text from original narratives. Li et al. (2021) evaluated the utility of 500 notes generated by text generation models for downstream entity recognition. The GReaT framework (Borisov et al., 2022) introduced a method for generating realistic tabular data by treating records as sequences, and Moser et al. (2024) proposed a multi-stage pipeline using large language models to generate structured patient-physician interactions in emergency medicine.

Current research focuses on high-fidelity generation and rigorous evaluation using large language models. Synthetic4Health (Ren et al., 2025) introduced a mask-and-generate framework for creating diverse, de-identified clinical letters. SynthMedic (Grazhdanski et al., 2025) proposed generating discharge summaries grounded in standard medical references (Merck Manuals) to ensure factual consistency without requiring access to real patient records. Meoni et al. (2025) framed synthetic text generation as an intermediate step for local model training, evaluating their approach on the MIMIC-III dataset. Sarkar et al. (2025) proposed a hybrid methodology, which combines de-identification with LLM filling to safely share clinical notes. Evaluation frameworks have also been standardized through toolkits like SynthTextEval (Ramesh et al., 2025), which assesses utility and privacy in health-related text. Finally, the MedGen model (Wang et al., 2025) scaled synthetic generation to multimodal domains, achieving strong performance in medical video generation.

These lines of work motivate documenting synthetic data provenance, grounding sources, and validation procedures, as such design choices influence the linguistic and clinical properties of the generated narratives. While the aforementioned literature relies heavily on automated generation, the synthetic narratives in our corpus were manually authored by domain specialists. This manual approach was chosen to strictly guarantee clinical fidelity, ensure adherence to specific diagnostic criteria without the risk of model hallucination, and provide a high-quality gold standard. In the corpus presented in this paper, these manually authored synthetic notes are integrated with non-synthetic material (adapted case reports and de-identified notes sampled from a public clinical corpus) to support analyses that distinguish between generated and naturally occurring documentation styles.

3 Corpus Design and Construction

This section describes the design principles and construction process of the proposed corpus. The dataset was created to support research on automatic detection of anaphylaxis in Portuguese clinical narratives, while addressing recurring challenges in clinical NLP, including data scarcity, class imbalance, and privacy constraints. To this end, we combined clinical texts from complementary sources (synthetic narratives, adapted case reports from the literature, and de-identified notes from a public clinical corpus), seeking to balance linguistic realism, ethical considerations, and reproducibility. In the context of this corpus, "synthetic" refers to fictional, representative clinical narratives

manually authored or edited by domain specialists (allergists) to simulate realistic patient encounters, rather than text generated automatically by software or language models. All texts, regardless of source, were subsequently reviewed and annotated by the specialists according to standardized clinical criteria. The following subsections detail the data sources, preprocessing and de-identification steps, and the annotation guidelines adopted.

3.1 Data Sources

The corpus comprises 969 short clinical narratives in Brazilian Portuguese, organized into three clinically motivated categories to support structured analysis and annotation consistency:

Synthetic cases (75 narratives). This subset consists of anonymized clinical narratives authored by domain specialists to mimic real-world medical records. From these, 35 notes have a confirmed diagnosis of anaphylaxis, and 40 were labeled as negative. These texts constitute the largest group of positive examples in the corpus.

Case reports from the literature (24 narratives). These narratives were adapted from published medical case reports. Thirteen texts describe confirmed cases of anaphylaxis, while eleven report other clinical conditions with overlapping or potentially confounding symptoms. All case reports were rewritten into a concise, note-like format to resemble electronic medical records while preserving clinically relevant information and removing any identifying details.

SemClinBr cases (870 narratives). The remaining texts were sampled from SemClinBr, a publicly available corpus of de-identified Brazilian Portuguese clinical narratives covering a wide range of medical conditions (Oliveira et al., 2022). To ensure sufficient informational content, only texts longer than 200 characters were selected. All SemClinBr narratives were independently reviewed by allergists; although allergic manifestations were observed in some cases, none satisfied the diagnostic criteria for anaphylaxis and were therefore labeled as negative.

Among the negative cases from the synthetic and case reports sources, 35 of them were considered as differential diagnosis cases. These are clinically challenging cases that present symptoms similar to anaphylaxis but correspond to alternative diagnoses. Ten narratives were adapted from published case reports, and twenty-five were derived from synthetic medical records. None of these cases met the diagnostic criteria for anaphylaxis, and all were labeled as negative after specialist review.

Overall, the corpus contains 48 positive cases of anaphylaxis and 921 negative cases. Positive cases originate exclusively from the synthetic and literature-derived subsets. The proportion of positive cases was deliberately fixed at approximately 5% to reflect prevalence estimates reported for emergency care settings (Cardona et al., 2020). All synthetic, anonymized, and adapted narratives were reviewed to ensure that they did not correspond to identifiable patient records. Table 1 presents representative excerpts from the corpus, sampled from each data source (English translations provided for convenience, dataset contains only Brazilian Portuguese text).

3.2 Preprocessing and De-identification

All texts were normalized using a preprocessing pipeline that removed formatting artifacts and corrected obvious spelling errors. Domain-specific terminology and common clinical abbreviations were preserved to maintain linguistic authenticity. No personally identifiable information (PII) was included in the synthetic or adapted notes. The SemClinBr texts were already de-identified in accordance with Brazilian data-protection regulations. We did not observe names, dates, or institutional identifiers in the final corpus. Given the intended public release of the dataset, these measures were deemed sufficient to mitigate privacy risks.

3.3 Annotation Guidelines

This subsection outlines the clinical criteria and procedures used to annotate the corpus for the presence or absence of anaphylaxis. The annotation guidelines were designed to operationalize established diagnostic definitions in a manner compatible with short, heterogeneous clinical narratives, enabling consistent labeling across all data sources while preserving clinical interpretability.

3.3.1 Clinical criteria for anaphylaxis

Annotation was guided by the NIAID/FAAN clinical criteria for anaphylaxis. A narrative was labeled as positive if at least one of the following conditions was satisfied (Cardona et al., 2020; Dribin et al., 2023):

- **Skin or mucosal involvement with respiratory or cardiovascular compromise:** Acute onset of urticaria, angioedema, or mucosal

Source	Label	Narrative text (Brazilian Portuguese)	English translation
Synthetic	Positive	Paciente refere prurido intenso pelo corpo, coçar, inchaços (pelotas) se formando no corpo (...) após sentir picadura/ferroada de inseto na perna (...)	Patient reports intense itching over the body, scratching, swellings (welts) forming on the body (...) after feeling an insect bite/sting on the leg (...)
Case report	Positive	Mal-estar geral, tontura, sudorese após ingesta de comprimido de omeprazol para dor epigástrica (...) observou-se edema da face e disartria (...)	General malaise, dizziness, sweating after ingestion of an omeprazole table for epigastric pain (...) facial edema and dysarthria were observed (...)
SemClinBr	Negative	Às 02:45h: encontra-se consciente (...) mantém acesso venoso periférico em MSE permeável salinizado no momento, apresenta rash cutâneo corporal + prurido (...)	At 02:45h: is conscious (...) maintains peripheral venous access in LUE [Left Upper Extremity], patent and saline-locked at the moment, presents with generalized skin rash + pruritus (...)
Differential diagnosis (synthetic)	Negative	Paciente (...) com quadro de edema de língua e lábios (deformante) com duração de 72h (...) PA 130/90mmHg, Sat O2 94%aa (...) Pele sem lesões (...)	Patient (...) with presentation of tongue and lip edema (deforming) lasting 72h (...) BP 130/90mmHg, O2 Sat 94% on room air (...) Skin without lesions (...)

Table 1: Excerpts from the corpus, sampled from each data source.

swelling accompanied by respiratory compromise (e.g., dyspnea, wheezing, stridor, hypoxemia) or by reduced blood pressure or syncope.

- **Multi-system involvement after exposure to a likely allergen:** Rapid involvement of at least two organ systems (mucocutaneous, respiratory, cardiovascular, or gastrointestinal) following exposure to a likely allergen, such as urticaria with vomiting or bronchospasm with hypotension.
- **Hypotension after exposure to a known allergen:** Sudden hypotension or syncope occurring after exposure to a known or highly probable allergen (Dribin et al., 2023).

The 2020 update by the World Allergy Organization recognizes that severe gastrointestinal symptoms or isolated respiratory or cardiovascular manifestations may also indicate anaphylaxis (Cardona et al., 2020). Annotators used these updates as supporting guidance but required that the core NIAID/FAAN criteria be met to assign a positive label. Negative labels were assigned when none of these criteria were satisfied or when the narrative clearly supported an alternative diagnosis.

3.3.2 Annotation procedure

Three board-certified allergists participated in annotating the 969 narratives, assigning a binary label indicating the presence or absence of anaphylaxis. Because the SemClinBr subset originates from a general clinical corpus with a low prior probability of anaphylaxis, each of these notes was reviewed by a single expert. Narratives from other sources were independently reviewed by at least two annotators. Annotators were provided only with the narrative text and did not have access to metadata regarding data source or patient context. Following independent annotation, cases of disagreement were discussed in consensus meetings. When disagreement persisted, a senior allergist acted as adjudicator.

3.3.3 Annotation format

Each entry in the released dataset contains the following fields:

- `note_id`: unique identifier (sequential integer);
- `text`: clinical narrative in Brazilian Portuguese;
- `source`: data source category (`synthetic`, `case_report`, or `semclinbr`);
- `label`: binary indicator of anaphylaxis presence (`1`) or absence (`0`);

The corpus is distributed as a single CSV file and may be converted to other structured formats, such as JSON, for convenience. For entries sourced from SemClinBr (source `semclinbr`), the `text` field contains the corresponding note ID rather than the full clinical narrative, as the original dataset requires credentialed access.

4 Corpus Analysis

This section characterizes the corpus in terms of basic composition and linguistic variability. We report high-level statistics on document length and lexical diversity, and describe how class labels and

writing styles differ across sources. The goal is to provide sufficient quantitative context for reproducibility while keeping the analysis focused on properties that affect downstream NLP use.

4.1 Document Length and Vocabulary

The corpus contains 969 clinical narratives with substantial variability in length, reflecting differences in writing conventions across sources. SemClinBr notes include both short chart-like entries and longer narratives, whereas synthetic notes tend to be more elaborated and explanatory. Using a simple regex-based tokenizer, the corpus contains approximately 129k tokens and 11.4k unique word forms, indicating a non-trivial lexical variety for a dataset of this size. Frequent terms and expressions are consistent with the clinical domain and include references to symptoms and interventions, e.g., “urticária” (hives), “edema” (edema), “dispneia” (dyspnea), “pressão arterial” (blood pressure), “epinefrina” (epinephrine), as well as common abbreviations (e.g., “PA” for blood pressure).

Table 2 summarizes the number of documents and class labels by source.

Source	Notes	Positive	Negative
Synthetic	75	35	40
Case reports	24	13	11
SemClinBr	870	0	870
Total	969	48	921

Table 2: Corpus composition by source and class label.

The “Synthetic” category aggregates expert-authored anaphylaxis cases and differential diagnosis cases derived from anonymized medical records.

4.2 Class Balance and Category Distribution

The dataset is deliberately imbalanced, with 48 positive cases of anaphylaxis (approximately 5%) and 921 negative cases, reflecting the rarity of the condition in routine clinical documentation (Cardona et al., 2020). Positive examples originate from the synthetic and literature-derived subsets, while SemClinBr contributes only negative examples. Importantly, all SemClinBr notes were manually reviewed by allergists, and although allergy-related content was observed in that subset, no cases meeting anaphylaxis criteria were identified. This design supports evaluation in settings where the target condition is rare and must be distinguished from both unrelated medical content and clinically similar presentations.

4.3 Linguistic Characteristics

The corpus exhibits systematic stylistic variation across sources, reflecting the different conditions of clinical documentation. De-identified notes from SemClinBr present a telegraphic structure, with an average sentence length of 9.8 words and a low lexical density of 0.26. This aligns with the prevalent use of institutional shorthand, abbreviations, and fragment syntax common in real-world charting, evidenced by an out-of-vocabulary (OOV) rate of 11.5%.

In contrast, adapted case reports and synthetic narratives present more complete, formal syntactic structures. Literature-derived reports average 23 words per sentence, while synthetic notes average 18.2. Both of these authored sources exhibit higher lexical density (approximately 0.46) and substantially lower OOV rates (1.7% and 2.5%, respectively). While the authored narratives explicitly describe termporal progression, triggers, and treatment responses, the SemClinBr notes compress this information into standard charting formats.

Explicit negation markers appear consistently across all subsets, ranging from 1.37 to 1.44 instances per 100 words. Positive cases are characterized by terms denoting temporal progressions and manifestations of the WAO criteria. Negative cases are weighted toward negation markers and vocabulary associated with alternative diagnoses. Standard clinical vocabulary such as "paciente" (patient) and "uso" (use) ranks highest across both classes. This heterogeneity makes the corpus suitable for studying domain adaptation and robustness to variation in clinical writing style.

5 Potential Applications

The annotated corpus enables a range of research applications in clinical natural language processing, particularly in scenarios involving rare-event detection and heterogeneous clinical narratives.

Document classification. The primary intended use of the corpus is the development and evaluation of models for automatic detection of anaphylaxis in free-text clinical notes.

The pronounced class imbalance reflects real-world conditions and makes the corpus suitable for studying methods for rare-event detection, cost-sensitive learning, and robustness under skewed

label distributions. Because the dataset mimics the low prevalence of rare clinical presentations, it is particularly well-suited for developing outlier detection algorithms, and evaluating the capabilities of pretrained models without the need for large-scale supervised training.

Information extraction. Beyond document-level classification, the corpus can be used to investigate the linguistic cues associated with anaphylaxis, including references to cutaneous manifestations, respiratory compromise, cardiovascular instability, and gastrointestinal symptoms. This supports the development of rule-based systems, feature-driven models, or explainable approaches that aim to align model decisions with established clinical criteria.

Language modeling and domain adaptation. The combination of synthetic narratives, adapted case reports, and de-identified real clinical notes provides a controlled setting for studying domain adaptation. Researchers may explore how language models trained on synthetic or literature-derived text generalize to authentic clinical documentation, and how mixed-source training affects performance in low-resource clinical domains.

Evaluation of large language models. Existing literature (Machado et al., 2024; Ensina et al., 2025) supports the application of this corpus as a benchmark for the evaluation of large language models, including GPT-3.5 and GPT-4. Possible extensions could systematically compare prompting strategies, zero-shot and few-shot settings, and domain-adapted models, in addition to analyzing error patterns in clinically challenging or ambiguous cases.

6 Conclusion

This paper introduced a dataset of 969 Brazilian Portuguese clinical narratives annotated for the presence or absence of anaphylaxis. The corpus integrates synthetic clinical notes, adapted case reports from the literature, and de-identified records from SemClinBr, with annotations grounded in established NIAID/FAAN and WAO diagnostic criteria. By combining heterogeneous text sources and explicitly documenting the annotation process, the dataset addresses a critical gap in publicly available clinical resources for Portuguese and supports research on rare-event detection in free-text medical narratives.

The detailed description of data provenance, labeling guidelines, corpus composition, and limitations aligns with current best practices for dataset transparency and responsible reuse. Rather than proposing new modeling approaches, this work aims to provide a well-characterized and reusable resource for the community. We expect the corpus to support reproducible studies in clinical NLP, including document classification, information extraction, and evaluation of language models, and to serve as a foundation for future extensions with richer annotations or additional clinical phenomena.

7 Limitations

Despite its contributions, the proposed corpus has limitations that should be considered when interpreting results obtained with it.

First, the dataset size is relatively modest, comprising 969 narratives. This limited number of instances presents a challenge for training traditional supervised machine learning models from scratch, as these models typically require larger corpora to achieve robustness without extensive data augmentation. However, this constraint reflects the fundamental challenge of sample size in rare disease and rare clinical event research (Schaefer et al., 2020). The dataset holds utility for alternative paradigms, such as outlier detection methods and pretrained models.

Second, positive cases account for approximately 5% of the dataset. This proportion reflects real-world prevalence estimates (Cardona et al., 2020) but results in a strongly imbalanced classification problem, which may adversely affect some learning algorithms if not explicitly addressed through appropriate evaluation protocols or training strategies.

Third, although synthetic notes were authored by experienced clinicians, they may not capture the full variability of language observed in real clinical documentation. Synthetic narratives may emphasize prototypical presentations of anaphylaxis and underrepresent atypical, incomplete, or ambiguously documented cases. Similarly, adapted case reports originate from published literature and may differ stylistically from routine electronic health records, which often contain fragmented or telegraphic language.

Finally, the corpus provides only a binary label indicating the presence or absence of anaphylaxis. It does not encode information about severity, trig-

gering agents, temporal progression, or treatment outcomes. Researchers interested in more fine-grained clinical modeling will need to extend the dataset with additional annotations.

8 Ethical Considerations

This section addresses ethical aspects related to the construction, annotation, and release of the corpus. Given the sensitive nature of clinical narratives, particular attention was paid to data privacy, anonymization, and responsible reuse. We also discuss potential biases introduced during data creation and annotation, as well as the intended research-only scope of the dataset.

8.1 Data Privacy and Anonymization

Clinical text frequently contains personally identifiable information, making data sharing ethically and legally challenging. To mitigate these risks, the corpus was constructed using a combination of synthetic data, adapted case reports, and de-identified clinical notes. Synthetic narratives are entirely fictional and do not correspond to real individuals. Adapted case reports were rewritten from published sources and stripped of identifying details. The SemClinBr subset is fully anonymised and complies with the Brazilian General Data Protection Law.

The use of synthetic data aligns with current recommendations advocating privacy-preserving alternatives for clinical NLP research (Mendes et al., 2025). Nevertheless, users of the corpus are encouraged to perform their own due diligence and ensure compliance with local regulations and institutional review requirements when integrating the dataset into downstream applications.

8.2 Annotation Guidelines and Clinical Responsibility

Annotations were performed using well-established diagnostic criteria (Cardona et al., 2020; Dribin et al., 2023), originally designed for clinical practice rather than text annotation. As a result, some degree of interpretative ambiguity is unavoidable, particularly when narratives provide incomplete symptom descriptions. Annotators relied on clinical judgment to resolve such cases, and disagreements were addressed through collective discussion.

The dataset is intended exclusively for research purposes. Models trained on this corpus should not be used for clinical decision-making without rigorous validation, regulatory approval, and integration into appropriate clinical workflows.

8.3 Bias and Representativeness

The corpus may reflect biases related to linguistic style, case selection, and data provenance. Synthetic notes were authored by a limited group of clinicians and may encode their preferred terminology or narrative structure. Adapted case reports originate from a subset of medical journals and may not represent the full diversity of healthcare settings or patient populations. The SemClinBr notes were collected from specific Brazilian institutions and may not generalize to other regions, languages, or health systems.

Researchers should therefore exercise caution when extrapolating findings beyond the scope of the dataset and consider complementing it with additional resources when addressing broader clinical or linguistic questions.

Acknowledgments

This work was supported by the University of São Paulo and the Sírio-Libanês College. This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brazil (CAPES) – Finance Code 001 and by grant CEPID 2013/07375-0 and grant C4AI 2019/07665-4 (C4AI), São Paulo Research Foundation (FAPESP).

References

Edmon Begoli, Kris Brown, Sudarshan Srinivas, and Suzanne Tamang. 2018. Synthnotes: A generator framework for high-volume, high-fidelity synthetic mental health notes. In *2018 IEEE international conference on big data (big data)*, pages 951–958. IEEE.

Emily M. Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Vadim Borisov, Kathrin Seßler, Tobias Leemann, Martin Pawelczyk, and Gjergji Kasneci. 2022. Language models are realistic tabular data generators. *arXiv preprint arXiv:2210.06280*.

Taxiarchis Botsis and R Ball. 2013. Automating case definitions using literature-based reasoning. *Applied clinical informatics*, 4(04):515–527.

Taxiarchis Botsis, Thomas Buttolph, Michael D Nguyen, Scott Winiecki, Emily Jane Woo, and Robert Ball. 2012. Vaccine adverse event text mining system for extracting features from vaccine safety reports. *Journal of the American Medical Informatics Association*, 19(6):1011–1018.

Victoria Cardona, Ignacio J Ansotegui, Motohiro Ebisawa, Yehia El-Gamal, Montserrat Fernandez Rivas, Stanley Fineman, Mario Geller, Alexei Gonzalez-Estrada, Paul A Greenberger, Mario Sanchez Borges, and 1 others. 2020. World allergy organization anaphylaxis guidance 2020. *World allergy organization journal*, 13(10):100472.

David S Carrell, Susan Gruber, James S Floyd, Maralyssa A Bann, Kara L Cushing-Haugen, Ron L Johnson, Vina Graham, David J Cronkite, Brian L Hazlehurst, Andrew H Felcher, and 1 others. 2023. Improving methods of identifying anaphylaxis for medical product safety surveillance using natural language processing and machine learning. *American Journal of Epidemiology*, 192(2):283–295.

Naila Camila da Rocha, Abner Macola Pacheco Barbosa, Yaron Oliveira Schnr, Juliana Machado-Rugolo, Luis Gustavo Modelli de Andrade, José Eduardo Corrente, and Liciana Vaz de Arruda Silveira. 2023. Natural language processing to extract information from portuguese-language medical records. *Data*, 8(1).

Rildo Pinto da Silva and Antonio Pazin-Filho. 2025. Dataset of anonymized discharge summaries of sepsis patients from a brazilian tertiary hospital for nlp applications. *Data in Brief*, page 111804.

Henrique Dias and Ana Helena Dias Pereira dos Ulbrich. 2022. BRATECA (Brazilian Tertiary Care Dataset): a Clinical Information Dataset for the Portuguese Language. PhysioNet. RRID:SCR_007345.

Timothy E Dribin, Megan S Motosue, and Ronna L Campbell. 2023. Overview of allergy and anaphylaxis. *Immunology and Allergy Clinics*, 43(3):435–451.

Luis Felipe Ensina, Matheus Matos Machado, Joice B. Machado Marques, Monica Pugliese H. dos Santos, Fábio Cerqueira Lario, Chayanne Andrade Araújo, Fabiana Andrade Nunes Oliveira, and Dilvan de Abreu Moreira. 2025. Artificial intelligence for detecting anaphylaxis in electronic medical records. *Asia Pacific Allergy*, 15(3):153–158.

Fernanda Bufon Färber, Iago Alves Brito, Julia Soares Dollis, Pedro Schindler Freire Brasil Ribeiro, Rafael Teixeira Sousa, and 1 others. 2025. Medpt: A massive medical question answering dataset for brazilian-portuguese speakers. *arXiv preprint arXiv:2511.11878*.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92.

Georgi Grazhdanski, Vasil Vasilev, Sylvia Vassileva, Dimitar Taskov, Izabel Antova, Ivan Koychev, and Svetla Boytcheva. 2025. Synthmedic: Utilizing large language models for synthetic discharge summary generation, correction and validation. *Journal of Biomedical Informatics*, 170:104906.

Alistair Johnson, Tom Pollard, Steven Horng, Leo Anthony Celi, and Roger Mark. 2023. MIMIC-IV-Note: Deidentified free-text clinical notes. *PhysioNet*. Version 2.2.

Jianfu Li, Yujia Zhou, Xiaoqian Jiang, Karthik Natarajan, Serguei Vs Pakhomov, Hongfang Liu, and Hua Xu. 2021. Are synthetic clinical notes useful for real natural language processing tasks: A case study on clinical entity recognition. *Journal of the American Medical Informatics Association*, 28(10):2193–2201.

Ying-Chih Lo, Sheril Varghese, Suzanne Blackley, Diane L Seger, Kimberly G Blumenthal, Foster R Goss, and Li Zhou. 2022. Reconciling allergy information in the electronic health record after a drug challenge using natural language processing. *Frontiers in allergy*, 3:904923.

Fábio Lopes, César Teixeira, and Hugo Gonçalo Oliveira. 2019. Contributions to clinical named entity recognition in portuguese. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 223–233.

Matheus Matos Machado, Joice Basílio Machado Marques, Fabrício A. Gualdani, Monica Pugliese Heleodoro dos Santos, Fabio Cerqueira Lario, Chayanne Andrade de Araujo, Fabiana Andrade Nunes Oliveira, Luis Felipe Chiaverini Ensina, Ricardo Marcondes Marcacini, and Dilvan Moreira. 2024. Evaluating large language models for anaphylaxis detection in clinical notes. *Journal of Health Informatics*, 16(Especial).

Jorge M Mendes, Aziz Barbar, and Marwa Refaie. 2025. Synthetic data generation: a privacy-preserving approach to accelerate rare disease research. *Frontiers in Digital Health*, 7:1563991.

Simon Meoni, Éric De La Clergerie, and Théo Ryffel. 2025. Synthetic documents for medical tasks: Bridging privacy with knowledge injection and reward mechanism. In *Proceedings of the Second Workshop on Patient-Oriented Language Processing (CL4Health)*, pages 12–25, Albuquerque, New Mexico. Association for Computational Linguistics.

Denis Moser, Matthias Bender, and Murat Sariyar. 2024. Generating synthetic healthcare dialogues in emergency medicine using large language models. In *Collaboration across Disciplines for the Health of People, Animals and Ecosystems*, pages 235–239. IOS Press.

Lucas Emanuel Silva Oliveira, Ana Carolina Peters, Adalniza Moura Pucca da Silva, Caroline Pilatti Gebeluca, Yohan Bonescki Gumiel, Lilian Mie Mukai Cintho, Deborah Ribeiro Carvalho, Sadid Al Hasan, and Claudia Maria Cabral Moro. 2022.

SemClinBr - a multi-institutional and multi-specialty semantically annotated corpus for portuguese clinical NLP tasks. *Journal of Biomedical Semantics*, 13(1).

Krithika Ramesh, Daniel Smolyak, Zihao Zhao, Nupoor Gandhi, Ritu Agarwal, Margrét V Bjarnadóttir, and Anjalie Field. 2025. Synthtexteval: Synthetic text data generation and evaluation for high-stakes domains. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 487–499.

Libo Ren, Samuel Belkadi, Lifeng Han, Warren Del-Pinto, and Goran Nenadic. 2025. Synthetic4health: generating annotated synthetic clinical letters. *Frontiers in Digital Health*, 7:1497130.

Atiquer Rahman Sarkar, Yao-Shun Chuang, Xiaoqian Jiang, and Noman Mohammed. 2025. Not fully synthetic: Llm-based hybrid approaches towards privacy-preserving clinical note sharing. *AMIA Summits on Translational Science Proceedings*, 2025:441.

Julia Schaefer, Moritz Lehne, Josef Schepers, Fabian Prasser, and Sylvia Thun. 2020. The use of machine learning in rare diseases: a scoping review. *Orphanet journal of rare diseases*, 15(1):145.

Mauricio Schiezaro, Guilherme Rosa, Bruno Augusto Goulart Campos, and Helio Pedrini. 2025. Guardians of the data: Ner and llms for effective medical record anonymization in brazilian portuguese. *Frontiers in Public Health*, 13:1717303.

Elisa Terumi Rubel Schneider, João Vitor Andrioli de Souza, Julien Knafou, Lucas Emanuel Silva e Oliveira, Jenny Copara, Yohan Bonescki Gumiel, Lucas Ferro Antunes de Oliveira, Emerson Cabrera Paraiso, Douglas Teodoro, and Cláudia Maria Cabral Moro Barra. 2020. BioBERTpt - a Portuguese neural language model for clinical named entity recognition. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, pages 65–72, Online. Association for Computational Linguistics.

Isabel Segura-Bedmar, Cristobal Colón-Ruíz, Miguél Ángel Tejedor-Alonso, and Mar Moro-Moro. 2018. Predicting of anaphylaxis in big data emr by exploring machine learning approaches. *Journal of biomedical informatics*, 87:50–59.

F Estelle R Simons, Ledit RF Ardusso, M Beatrice Bilò, Yehia M El-Gamal, Dennis K Ledford, Johannes Ring, Mario Sanchez-Borges, Gian Enrico Senna, Aziz Sheikh, Bernard Y Thong, and 1 others. 2011. World allergy organization guidelines for the assessment and management of anaphylaxis. *World Allergy Organization Journal*, 4(2):13–37.

Kathleen E Walsh, Sarah L Cutrona, Sarah Foy, Meghan A Baker, Susan Forrow, Azadeh Shoaibi, Pamala A Pawloski, Michelle Conroy, Andrew M Fine, Lise E Nigrovic, and 1 others. 2013. Validation of anaphylaxis in the food and drug administration's mini-sentinel. *Pharmacoepidemiology and drug safety*, 22(11):1205–1213.

Rongsheng Wang, Junying Chen, Ke Ji, Zhenyang Cai, Shunian Chen, Yunjin Yang, and Benyou Wang. 2025. Medgen: Unlocking medical video generation by scaling granularly-annotated medical videos. *arXiv preprint arXiv:2507.05675*.

Yanshan Wang, Sunyang Fu, Feichen Shen, Sam Henry, Ozlem Uzuner, and Hongfang Liu. 2020. The 2019 n2c2/ohnlp track on clinical semantic textual similarity: Overview. *JMIR Med Inform*, 8(11):e23375.

Wei Yu, Chengyi Zheng, Fagen Xie, Wansu Chen, Cheryl Mercado, Lina S Sy, Lei Qian, Sungching Glenn, Hung F Tseng, Gina Lee, and 1 others. 2020. The use of natural language processing to identify vaccine-related anaphylaxis at five health care systems in the vaccine safety datalink. *Pharmacoepidemiology and drug safety*, 29(2):182–188.

LLM-Based Multi-Agent System with Retrieval-Augmented Generation for Medical Care Planning Generation in Sickle Cell Disease

Luana Bringel Leite[1], David Eduardo Pereira[1], Eyshila Buriti de Araujo Azevedo[1], Leonardo Mota Meira Filho[1], Eliane Cristina Araújo[1], Claudio E. C. Campelo[1], Taciana R. O. C. Marques [2], Letícia B. de Almeida[2], Herman Martins Gomes[1]

[1]Federal University of Campina Grande (UFCG) - Academic Unity of Systems and Computing
[2] Federal University of Campina Grande (UFCG) - Center for Biological and Health Sciences

Correspondence: luana.leite@ccc.ufcg.edu.br

Abstract

Ensuring safety in clinical applications of large language models (LLMs) remains an unresolved challenge, particularly for high-risk and underrepresented conditions such as Sickle Cell Disease (SCD). Consequently, these models may exhibit limited reliability for SCD, including hallucinations and clinical non-adherence. This paper proposes an LLM-based Multi-Agent System (MAS) enhanced by Retrieval-Augmented Generation (RAG) to automate the generation of medical care plans for SCD. The MAS decomposes clinical reasoning into specialized agents responsible for diagnosis, investigation, and treatment planning. Retrieval is framed not as a performance optimization, but as a safety control mechanism. Three RAG strategies, namely LLM-Guided Tree Retrieval, Metadata-Filtered Retrieval, and Semantic Similarity Retrieval, are evaluated alongside a baseline. Our experiments considered LLM-as-a-Judge evaluations and independent assessments by physicians. The results demonstrate high clinical quality, with safety scores exceeding 4 on a 5-point scale. While average performance was similar between RAG and baseline conditions, the Tree Retrieval strategy reduced the frequency of clinically unsafe outputs compared to conventional Semantic Retrieval, indicating fewer clinically unsafe outputs. These findings provide evidence that average performance is insufficient to evaluate clinical AI systems, particularly in high-risk scenarios where retrieval serves as a safety control layer.

1 Introduction

Creating a medical care plan for a patient is a critical task for healthcare professionals. A care plan has a direct impact on a patient's life quality (Abdeldafie and Alaajmi, 2022), as it guides clinical procedures, medication management, diagnostic examinations, monitoring protocols, among other interventions. This process becomes more challenging when designing care plans for patients with rare diseases or conditions, a scenario that physicians rarely encounter during medical education or routine practice. (Walkowiak and Domaradzki, 2021).

Sickle Cell Disease (SCD) is a rare genetic disorder that, according to existing studies, will affect around 400,000 individuals worldwide until 2050, including a significant number in Brazil (estimate of 30,000) (Kato et al., 2018). Given its prevalence in specific populations, SCD is often underrepresented in medical education and insufficiently understood by physicians, leading to inadequate management in clinical practice (Druye et al., 2024). This lack of familiarity is further influenced by broader social, cultural, and historical factors, since SCD disproportionately affects individuals of African descent due to genetic inheritance.

Aligned with this social and cultural marginalization, significant knowledge gaps persist among physicians regarding SCD (Druye et al., 2024). Structural inequalities have contributed to its underrepresentation in medical education, research and clinical practice (Reich et al., 2022). This issue extends to technological development within the field of computer science and healthcare informatics. Studies indicate that SCD, along with other rare and African diseases, is underrepresented in healthcare evaluation datasets (Mutisya et al., 2025). As a result, AI models may exhibit limited performance and reduced reliability when addressing these conditions, reinforcing gaps in access to specialized care.

This research proposes an LLM-based Multi-Agent System (MAS) enhanced by Retrieval-Augmented Generation (RAG) for automated medical care plan generation in SCD as a module that is part of *HemaChat*[1] software, a multi-agent clinical reasoning and decision support system designed to expand access to safe medical guidance for SCD

[1] https://bit.ly/hemachat

Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)
Vol. 2, pages 88–100

patients. The proposed MAS explicitly decomposes clinical reasoning into sequential specialized agents. By integrating structured retrieval at each reasoning stage, the system constrains generation within validated clinical protocols, improving reliability, interpretability, and clinical safety. Previous studies indicate that RAG can significantly improve the quality and reliability of LLM-based applications in healthcare (Amugongo et al., 2025). In safety-critical domains, retrieval architecture should be understood as a safety control layer rather than a performance optimization, as it directly constrains generation and reduces the likelihood of harmful outputs.

This research evaluates different types of RAG techniques in the context of medical care plan generation for SCD on patient-reported symptoms. Since existing models can demonstrate limited capabilities when dealing with rare diseases, the application of RAG can enhance reliability and performance in medical LLM-based systems. Therefore, the present study evaluates three distinct RAG techniques (Semantic Similarity Retrieval, Metadata-Filtered Retrieval, and LLM-Guided Tree Retrieval), incorporating both **LLM-as-a-judge** (Li et al., 2025) and **human** evaluations conducted by physicians.

This study offers three key contributions: (1) a structured multi-agent architecture aligned with clinical reasoning workflows; (2) an evaluation of retrieval mechanisms for safety-critical clinical applications; and (3) empirical evidence supporting retrieval as a safety control layer in generative AI systems, demonstrating its role in reducing unsafe outputs in high-risk clinical scenarios.

The remainder of this paper is structured as follows. Section 2 reviews related work on LLMs and Retrieval-Augmented Generation in healthcare. Section 3 presents the proposed multi-agent architecture, the RAG strategies, and the clinical evaluation protocol. Section 4 reports the experimental results. Finally, Section 5 discusses the implications, limitations, and future research directions.

2 Related Work

The use of LLMs in healthcare is a rapidly expanding research field (Wang et al., 2024b). AI systems have demonstrated the ability to support a wide range of healthcare-related tasks. Existing review studies show that LLM applications in healthcare are broad, encompassing the summarization of complex clinical information, medical knowledge retrieval to support question answering and examinations, improved public access to medical information, predictive tasks such as diagnosis, treatment support, and drug interaction analysis, as well as administrative activities, including clinical documentation and public health data collection (Wang et al., 2024b).

Despite these broad applications, the use of LLMs in healthcare raises several ethical concerns that must be addressed to prevent harm (Wang et al., 2023). These concerns span legal, humanistic, algorithmic, and informational dimensions, including unclear liability in cases of patient harm, risks to patient privacy, potential disruption of the physician–patient relationship, erosion of trust due to over-reliance on AI, and challenges related to transparency, bias, and explainability (Wang et al., 2023).

Recent advances in LLMs demonstrate improved reasoning capabilities, reduced latency, and multimodal functionality, which help mitigate some challenges in healthcare applications (Neha et al., 2025). Notably, many LLMs have been fine-tuned on biomedical corpora to enhance domain-specific comprehension (Neha et al., 2025), including recent ChatGPT variants specifically designed for healthcare-related queries[2]. However, these models cannot continuously incorporate evolving clinical knowledge, which limits their adaptability in dynamic healthcare environments and are susceptible to hallucinations.

To address these limitations, RAG techniques have emerged as a promising approach to improve the reliability of LLMs. RAG helps mitigate hallucinations and reduces over-reliance on static model training data (Arslan et al., 2024). Nevertheless, RAG is not a silver bullet. Studies indicate that hallucinations can still occur and that factual inconsistencies remain a persistent issue even in RAG-based systems applied to healthcare scenarios (Amugongo et al., 2025).

The study proposed by Neha et al. (2025) emphasizes the use of RAG in healthcare domains such as diagnostic assistance, electronic health record and discharge note summarization, medical question answering, patient education and conversational agents, clinical trial matching, and biomedical literature synthesis. Beyond these areas, research

[2]https://openai.com/pt-BR/index/introducing-chatgpt-health/

has explored the application of RAG in more specialized healthcare domains. These include mental health–related solutions (Kermani et al., 2025), patient simulation for educational purposes (Yu et al., 2025), health problem identification in home healthcare settings (Zhang et al., 2025), inclusive urban public healthcare services (Sun et al., 2025), among other context-specific applications.

It is important to highlight that RAG can be a valuable tool for mitigating hallucinations and other LLM-related issues in healthcare scenarios. However, studies also reveal significant limitations. Most research focuses on English and Chinese, leaving many other languages underrepresented (Amugongo et al., 2025). Moreover, bias can persist or even be reproduced through RAG pipelines, as biased source data can propagate biased outputs, potentially leading to harmful or misleading information. Another limitation is that current evaluation metrics are often insufficient to assess RAG performance in healthcare contexts, as they may not adequately capture clinical relevance or safety considerations (Neha et al., 2025).

For these reasons, this research investigates the effectiveness of LLM-based MAS for medical care plan generation in SCD, enhanced by RAG techniques. It combines quantitative metrics with assessments generated by specialized human and LLM evaluations. Furthermore, the focus on SCD provides an important contribution, given the limitations of LLM in rare disease contexts that are often underrepresented in datasets and LLM models (Mutisya et al., 2025). This work also takes into account the Brazilian Portuguese language setting, addressing another critical gap in current healthcare-focused LLM research, which is predominantly centered on the English language.

It is important to note that research on rare diseases, such as SCD, remains limited and often receives insufficient attention, mainly due to the relatively small number of affected individuals (Visibelli et al., 2023). When surveying the literature on SCD and AI, only a small number of relevant studies can be identified. These include research on the use of LLMs in ambulatory devices for home health diagnostics, a case study focused on sickle cell anemia management (Ogundare and Sofolahan, 2023). Additionally, a question-answering study addresses rare diseases more broadly rather than exclusively focusing on SCD (Wang et al., 2024a).

Furthermore, when examining research on medical care plan generation, a similar scarcity of studies is observed. One notable example is MEDPlan (Hsu et al., 2025), an approach for medical care plan generation that leverages LLMs and RAG within the Subjective, Objective, Assessment, Plan (SOAP) framework, however, it does not directly address the limitations of the SCD context. To the best of our knowledge, there is currently no research investigating the use of LLM-based MAS enhanced by RAG in the context of the generation of medical care plans specifically tailored to SCD.

3 Methodology

This study is a prospective, blinded clinical validation to evaluate the adequacy, completeness, and safety of medical care plans generated by an LLM-based MAS enhanced with RAG within a broader multi-agent clinical reasoning and decision support system. All clinical cases, system outputs, and evaluation procedures were predefined prior to assessment, and physician evaluators were blinded to the retrieval strategy, ensuring unbiased assessment.

The system itself is intentionally engineered to replicate the sequential reasoning process employed by physicians in emergency care settings as a decision support tool (Croskerry, 2009). In real clinical workflows, physicians do not generate diagnoses, investigations, and treatments simultaneously; rather, they follow a *structured reasoning sequence* in which initial clinical observations inform diagnostic hypotheses, which in turn guide selection of confirmatory investigations and ultimately determine therapeutic interventions. This process is iterative and uncertainty-aware, and is mirrored by the proposed MAS system (see Section 3.3 for details).

The experimental protocol compared three RAG strategies: LLM-Guided Tree Retrieval (proposed method), Metadata-Filtered Retrieval, and Semantic Similarity Retrieval, each evaluated against a baseline without retrieval. Each generated medical care plan was independently evaluated by three blinded physicians and, in parallel, by an automated LLM-as-a-Judge framework (Gu et al., 2024). This dual evaluation design allowed for a direct comparison between human clinical judgement and automated evaluation methods.

3.1 Medical Care Plan Data

To ensure validity and faithful representation of real-world clinical complexity, 10 clinical case vi-

gnettes were prospectively developed by a senior pediatric and SCD specialist. These cases were specifically designed to represent the spectrum of acute SCD emergencies most frequently encountered in emergency departments, including vaso-occlusive crisis, acute chest syndrome, splenic sequestration, ischemic priapism, and acute neurological complications such as ischemic stroke and seizure presentations, which together represent the most common and clinically significant acute manifestations of SCD (Rees et al., 2010; Piel et al., 2017).

These scenarios represent high-risk emergency conditions that require timely and appropriate management, making them suitable for the evaluation of critical safety systems. Although the number of clinical cases included in this study is limited, this study prioritizes depth of clinical validation over scale. Each vignette represents a high-risk emergency scenario and was carefully designed by a specialist to reflect real-world complexity and clinical decision-making requirements and cover most recurrent acute SCD complications.

In addition, each vignette was constructed using a realistic emergency department triage structure. A specialist physician generated a comprehensive gold-standard medical care plan representing the clinically optimal management approach for each patient. The standardized clinical structure comprised patient identification and clinical context, clinical history and presenting complaint, physical examination findings, differential diagnosis, diagnostic investigations, therapeutic management, and follow-up monitoring. A detailed description and a complete example of the vignettes are provided in Figure 4 in the Appendix A.6 for illustration.

These reference medical care plans were constructed using current institutional protocols and established clinical guidelines and served as the *clinical reference benchmark* against which AI-generated medical care plans were evaluated. Importantly, the Gold Standard was not directly provided to the AI system during generation, ensuring that evaluation reflected true generalization rather than memorization.

3.2 RAG Knowledge Base

To ensure clinical safety and prevent the generation of recommendations based on unreliable or unverified sources, the knowledge base of the system was constructed exclusively from validated institutional clinical protocols and professional society guidelines.

The knowledge corpus consists of:

1. **The Pediatric Emergency Care Protocol** of the *Hospital de Clínicas de Porto Alegre* (HCPA) (Hospital de Clínicas de Porto Alegre, 2023), and

2. **Clinical management guidelines** from the *Sociedade de Pediatria do Estado do Rio de Janeiro* (SOPERJ) (Sociedade de Pediatria do Estado do Rio de Janeiro, 2023).

These sources are authoritative clinical references used in pediatric emergency care in Brazil. Restricting the knowledge base to curated clinical protocols is a critical safety design decision, as open-web retrieval can introduce outdated, contradictory, or non-validated medical information (Institute of Medicine, 2011).

Documents were segmented into semantically coherent text fragments (*"chunks"*) and indexed within a ChromaDB vector database[3] using all-MiniLM-L6-v2 embeddings[4]. This embedding model was selected due to its performance in semantic retrieval tasks while maintaining computational efficiency.

The retrieval hyperparameter was set to top-k = 5, meaning that the five most relevant knowledge fragments were retrieved for each query. This choice is consistent with prior RAG research, which shows that retrieval configuration and document relevance influence reasoning reliability and hallucinations (Lewis et al., 2020; Yan et al., 2024). Retrieving too few documents may lead to incomplete clinical context, whereas retrieving too many may introduce irrelevant information that degrades generation quality.

3.3 Implementation Details

All agents and experimental conditions were powered by the GPT-4.1 model[5] via the OpenAI API, and agent communication was implemented using the Python-based LangChain[6] framework. For safety-critical reasoning tasks, generation was performed with the temperature set to 0, promoting

[3] https://www.trychroma.com/

[4] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

[5] https://developers.openai.com/api/docs/models/gpt-4.1

[6] https://www.langchain.com/

deterministic and stable outputs; this configuration reduces stochastic variability and improves reproducibility in clinical decision-making scenarios. Furthermore, all retrieval strategies and the baseline used the same underlying LLM, ensuring that observed differences are attributable to retrieval mechanisms rather than to model variation.

3.4 Retrieval Strategies

This study investigates three different retrieval approaches, each with its own data representation and retrieval mechanism, as described below.

Semantic Similarity Retrieval: This standard dense approach retrieves protocol fragments using dense vector similarity. Queries and document chunks are embedded using the all-MiniLM-L6-v2, and the top-k fragments are selected using Maximal Marginal Relevance (MMR).

Metadata-Filtered Retrieval: This strategy extends semantic retrieval by applying metadata filters before similarity search. Protocol fragments are first restricted based on attributes such as disease and protocol category, and similarity search is performed within this subset.

LLM-Guided Tree Retrieval: This strategy implements a structured LLM-guided retrieval mechanism over a hierarchical representation of clinical knowledge. The knowledge base is organized as a tree structure, where each node represents a clinically meaningful unit (e.g., symptoms, diagnostic categories, or treatment protocols). Each node contains a unique identifier, a title summarizing its clinical meaning, a short description, and the associated clinical content.

Given a clinical vignette, the LLM analyzes the patient symptoms and selects the most relevant nodes from the tree. This selection is performed by providing the model with a structured representation of the tree (excluding full clinical text) and prompting it to identify relevant node identifiers. Hence, the retrieval process consists of two stages: (1) node selection, in which the LLM identifies relevant nodes, and (2) content extraction, in which the full clinical content associated with these nodes is retrieved and concatenated to form the final context.

Unlike conventional semantic retrieval, which operates on flat document representations, the tree approach constrains retrieval to clinically coherent pathways, reducing the likelihood of retrieving contextually irrelevant but lexically similar information. Additionally, the model produces an intermediate reasoning trace during node selection, enabling interpretability of retrieval decisions.

3.5 LLM-Based Multi-Agent System Architecture

The proposed system utilizes a sequential LLM-Based Multi-Agent architecture in which each agent performs a specialized stage of the clinical reasoning process, thereby improving interpretability and enabling fine-grained safety analysis. Rather than generating medical care plans in a single step, the system progresses through a structured reasoning pipeline, as illustrated in Figure 1, which presents the complete multi-agent clinical reasoning pipeline, listed in the following paragraphs:

Diagnostic Hypothesis Agent: This agent initiates the clinical reasoning process by analyzing the clinical vignette and retrieving protocol-grounded knowledge to construct a structured differential diagnosis. Each hypothesis is explicitly categorized as *Most probable*, *Less probable*, *To be ruled out*, or *Diagnosis of exclusion*, enabling formal representation of clinical uncertainty and systematic prioritization of high-risk conditions.

Diagnostic Investigation Agent: Building upon the diagnostic hypotheses, this agent retrieves protocol-aligned recommendations for diagnostic investigations. By conditioning test selection on explicit reasoning outputs, the system ensures clinical coherence and reduces the risk of incomplete or unjustified evaluations.

Treatment Agent: This agent generates therapeutic recommendations grounded in validated clinical protocols, taking into account the prioritized diagnostic hypotheses and clinical severity. This structured grounding constrains generation within clinically accepted standards and improves treatment safety.

Medical Care Plan Generation Agent: The final agent integrates all intermediate outputs into a unified and structured medical care plan. This modular synthesis preserves traceability across reasoning stages and produces coherent clinically actionable documentation aligned with real-world clinical workflows.

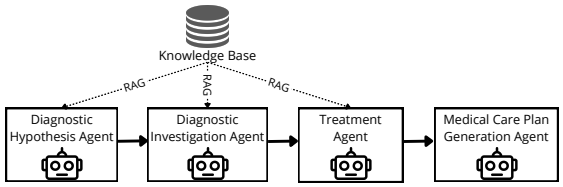


Figure 1: Overview of the proposed LLM-based Multi-Agent Retrieval-Augmented Generation architecture for medical care plan generation.

3.6 Human Clinical Evaluation Protocol

The generated medical care plans were evaluated through a prospective, blinded physician assessment protocol. A total of 22 physicians participated as evaluators, including both board-certified specialists and physicians in residency training. This mixed-expertise composition reflects clinicians who may interact with AI-generated medical care plans in real healthcare settings, including those with limited access to specialized SCD expertise.

The experimental design comprised 40 distinct AI-generated medical care plan documents, corresponding to 10 clinical cases processed under each of the three retrieval strategies and the baseline. Each medical care plan was independently evaluated by exactly 3 different physicians, resulting in a total of 120 blinded physician evaluations as summarized in Figure 2 (Appendix A.1).

To preserve blinding and prevent expectation bias, physicians were not informed of the generation method, RAG strategy, or experimental condition associated with any medical care plans. For each clinical vignette, physicians evaluated a set of medical care plans that included both the gold-standard plan developed by a specialist and those generated by the experimental systems, presented in random order and without source identification. Each plan was presented in conjunction with the original clinical vignette and assessed solely on its clinical merits, consistent with routine clinical decision-making.

Physicians independently assessed each medical care plan using a **standardized 5-point Likert scale (1–5)** across three clinically critical evaluation criteria: Clinical Adequacy, Completeness, and Clinical Safety, as defined in Appendix A.3. In addition, the evaluators reported their self-assessed expertise in SCD to characterize the evaluation cohort. Two self-reported measures were used: **(1) general knowledge of the disease** and **(2) knowledge of its clinical management and treatment**. Participants rated their experience using a 5-point Likert scale ranging from Level 1 (Very low) to Level 5 (Very high). The results are described in Table 3 in Appendix A.5.

3.6.1 Evaluation Distribution and Randomization

Medical care plans were assigned using a custom Python-based constrained randomization algorithm. Each plan was independently evaluated by exactly three physicians to ensure reliability. Workload was balanced to minimize fatigue effects: of the 22 participating physicians, 10 completed six evaluations and 12 completed five evaluations. All assignments were randomized and blinded with respect to the generation method and RAG strategy. Physicians evaluated each plan based solely on its clinical content. This ensured a balanced and unbiased evaluation dataset.

3.7 Automated LLM-as-a-Judge Evaluation

In parallel with human evaluation, all generated medical care plans were independently evaluated using an **automated LLM-as-a-Judge (Croxford et al., 2025) framework** based on GPT-4o-mini[7]. This automated evaluation used identical evaluation criteria and scoring scales as the human evaluators

The **LLM-Judge** was provided with the same clinical vignette and generated medical care plans as input (the prompt is provided in Figure 3 in Appendix A.2), and its scoring was performed in isolation without access to human ratings or experimental condition information. This parallel evaluation design enabled direct, paired comparison between human expert assessment and automated evaluation, allowing systematic analysis of agreement, bias, and safety detection performance between human and AI evaluators.

3.8 Statistical Analysis

Statistical analysis was designed to evaluate both **overall performance trends and clinically critical tail-risk safety behavior**. In clinical safety research, average performance alone is insufficient, as patient harm is typically driven by low-probability but high-impact unsafe outputs. Accordingly, the statistical framework incorporated both *central tendency and risk-focused safety analyses*.

[7] https://platform.openai.com/docs/models/gpt-4o-mini

Descriptive statistics were computed for each retrieval strategy and evaluation criteria, with results summarized as mean and standard deviation. To assess the stability and precision of these estimates, 95% confidence intervals were computed using bootstrap resampling. Although Likert data are ordinal, means and standard deviations are reported for interpretability and comparability, while inferential analyses were conducted using non-parametric tests.

To evaluate differences between retrieval strategies, non-parametric group comparisons were performed using the **Kruskal–Wallis test**, appropriate for ordinal Likert-scale data without assuming normal distribution. When the Kruskal–Wallis test was applied, pairwise comparisons were conducted using **Dunn's post-hoc test** with **Holm correction** for multiple comparisons. For the direct comparison between human and LLM-Judge evaluations, where scores were paired per medical care plan, the **Wilcoxon signed-rank test** was used. Effect sizes were additionally quantified using **Cohen's** d for pairwise comparisons and **epsilon-squared** (ε^2) for Kruskal–Wallis tests, providing standardized measures of effect magnitude independent of statistical significance.

Critically, to directly assess clinical safety risk, a Low-Safety Rate analysis was performed, defined as the proportion of medical care plans that received a Clinical Safety score of 3 or lower. This metric captures the frequency of clinically concerning outputs and provides a direct measure of patient safety risk exposure.

Finally, to quantify agreement between human and automated evaluation, the Mean Absolute Error (MAE) was calculated between human and LLM-Judge scores. This comprehensive statistical framework enabled robust evaluation of both average performance and clinically meaningful safety risk. Furthermore, a two-sided significance level of $\alpha = 0.05$ was used for all statistical tests.

3.9 Inter-Rater Reliability

To assess the reliability and consistency of physician evaluations, inter-rater agreement was quantified using **Krippendorff's Alpha**, a statistical measure appropriate for ordinal data and multiple independent raters (Krippendorff, 2011). The values (presented in Table 1) indicate **moderate variability** among evaluators, reflecting the inherent complexity and subjective nature of clinical judgment in emergency management of SCD.

Table 1: Inter-rater agreement among physician evaluators measured using Krippendorff's Alpha

Evaluation Criteria	Krippendorff's Alpha
Clinical Adequacy	0.46
Completeness	0.45
Clinical Safety	0.45

Importantly, as discussed earlier, the evaluation protocol incorporated **triple redundancy**, with each medical care plan independently assessed by three physicians. This design improves the reliability of aggregated scores and reduces the influence of individual evaluator variability. As a result, the reported findings reflect stable **collective clinical judgment** rather than isolated subjective opinions, supporting the robustness of the human evaluation framework.

4 Results

This section reports clinical safety outcomes for the evaluated LLM approaches based on a blinded physician assessment. It presents comparative statistical analysis of mean safety scores, a tail-risk evaluation of rare high-impact unsafe outputs. The section concludes with a comparison between physician evaluations and LLM-as-a-Judge assessments.

4.1 Clinical Safety Performance Based on Human Physician Evaluation

LLM-Guided Tree Retrieval achieved the highest overall Clinical Safety performance among the evaluated strategies based on physician ratings. Mean Clinical Safety scores were 4.10 (SD = 0.88) for LLM-Guided Tree Retrieval, compared to 4.07 (SD = 0.64) for the Baseline, 4.03 (SD = 0.81) for Metadata-Filtered Retrieval, and 3.77 (SD = 1.04) for Semantic Similarity Retrieval.

The absolute improvement of 0.33 points in Clinical Safety between LLM-Guided Tree Retrieval and Semantic Similarity Retrieval corresponds to a small-to-moderate effect size (Cohen's $d = 0.34$), indicating a meaningful reduction in safety risk magnitude. LLM-Guided Tree Retrieval and Baseline exhibited comparable mean performance, differing by only 0.03 points. However, qualitative analysis of physician comments (see Appendix A.4) revealed that baseline-generated plans more frequently contained missing critical steps, insufficient justification of diagnostic reasoning, and

occasional unsafe omissions, indicating that similar mean scores may obscure clinically relevant safety deficiencies.

The confidence interval analysis further demonstrated improved lower-bound safety performance for LLM-Guided Tree Retrieval. LLM-Guided Tree Retrieval achieved a mean Clinical Safety score of 4.10 (95% CI: 3.77–4.43), whereas Semantic Similarity Retrieval achieved 3.77 (95% CI: 3.38–4.16). In particular, the lower bound of LLM-Guided Tree Retrieval performance approached the mean safety level of Semantic Similarity Retrieval, suggesting improved worst-case safety performance.

These findings highlight that average performance alone is insufficient for evaluating clinical AI systems in safety-critical settings. Although mean scores were similar across strategies, safety-critical differences emerged in tail-risk analysis, indicating that retrieval design plays a crucial role in reducing clinically unsafe outputs. Importantly, systems with comparable average performance may exhibit substantially different safety profiles, indicating that mean-based evaluation can mask clinically significant risks.

4.2 Comparative Statistical Analysis

Despite the observed differences in mean Clinical Safety scores, the Kruskal-Wallis analysis did not detect statistically significant differences between the retrieval strategies (H = 1.47, $p = 0.69$). This lack of statistical significance may reflect the limited sample size and relatively high baseline performance across conditions.

However, effect size analysis indicated significant practical differences between retrieval architectures, particularly between LLM-Guided Tree Retrieval and Semantic Similarity Retrieval (Cohen's $d = 0.34$), supporting the presence of clinically relevant safety improvements not fully captured by null hypothesis testing alone.

4.3 Tail-Risk and Unsafe Output Reduction

Tail-risk analysis based on physician evaluation revealed substantial differences in the frequency of clinically unsafe outputs. Unsafe outputs were operationally defined as medical care plans receiving Clinical Safety scores ≤ 3, representing recommendations considered potentially unsafe or clinically concerning.

Semantic Similarity Retrieval produced unsafe outputs in 30.0% of cases. In contrast, LLM-Guided Tree Retrieval reduced this proportion to 13.3%, representing a 55.6% relative reduction in unsafe outputs.

This represents a clinically meaningful improvement, as patient harm is driven by rare high-severity failures rather than average performance alone. The observed reduction in unsafe recommendations suggests that structured retrieval substantially improves the safety risk profile of generated medical care plans. Importantly, physician feedback indicated that unsafe baseline outputs often resulted from incomplete clinical reasoning chains and lack of protocol grounding, reinforcing that baseline generation may produce superficially adequate but clinically fragile plans.

4.4 Safety Variability and Reliability

Semantic Similarity Retrieval exhibited substantially higher variability in Clinical Safety scores (SD = 1.04) compared to LLM-Guided Tree Retrieval (SD = 0.88). This increased variance indicates greater unpredictability in system performance and a greater likelihood of safety failures.

From a safety engineering perspective, reduced variance represents improved system reliability and greater consistency in clinical output quality. Structured retrieval constrains information access to clinically coherent protocol pathways, reducing the risk of contextually inappropriate retrieval and improving output stability. Hence, this improved reliability is critical for clinical deployment, where unpredictable behavior may increase patient risk.

4.5 Comparison Between Human and LLM-as-a-Judge Evaluation

Comparison between blinded human physician evaluation and automated LLM-as-a-Judge assessment revealed substantial and statistically significant discrepancies in safety perception.

Across all 120 evaluations, human physicians assigned a mean Clinical Safety score of 3.99, whereas the automated LLM-Judge assigned a substantially higher mean score of 4.93, representing an absolute difference of 0.94 points (23.5% of the Likert scale range). This difference was statistically significant (Wilcoxon signed-rank test, $p = 3.22 \times 10^{-15}$).

The magnitude of disagreement was further quantified using Mean Absolute Error (MAE), which was 1.00 across all evaluations, indicating that automated safety ratings deviated by approximately one full Likert point on average (Figure 5,

Appendix A.7). This discrepancy was systematic rather than random. The LLM-Judge consistently overestimated safety across all retrieval strategies, assigning near-ceiling scores even in cases where physicians identified clinically meaningful safety concerns. Notably, several baseline plans rated as safe by the LLM-Judge were flagged by physicians as clinically incomplete or potentially unsafe, further highlighting that mean-based automated evaluation may fail to detect critical safety issues.

Strategy-specific analysis demonstrated this pattern consistently (Table 2). Notably, Semantic Similarity Retrieval received a perfect mean safety score of 5.00 from the LLM-Judge despite receiving the lowest safety ratings from human physicians. These findings show that automated evaluation systematically fails to detect clinically meaningful safety risks, highlighting the **necessity of human expert evaluation**.

Table 2: Human vs LLM-Judge safety scores (mean ± SD) and MAE

Strategy	Human	LLM	MAE
LLM-Guided Tree	4.10 ± 0.88	4.90 ± 0.31	0.87
Metadata-Filtered	4.03 ± 0.81	4.90 ± 0.31	1.00
Semantic Similarity	3.77 ± 1.04	5.00 ± 0.00	1.23
Baseline	4.07 ± 0.64	4.90 ± 0.31	0.90

5 Conclusion and Future Work

This study presents a clinically validated LLM-based MAS with RAG, developed within the HemaChat[8] clinical reasoning and decision support system, capable of generating safe and high-quality medical care plans for SCD enabling broader access to safe clinical guidance. By decomposing clinical reasoning into specialized agents grounded in validated protocols, the system enables reliable clinical document generation with LLMs.

Furthermore, through a prospective, blinded evaluation involving 22 physicians and 120 independent clinical assessments, the proposed LLM-Guided Tree Retrieval architecture achieved high clinical adequacy, completeness, and safety, while reducing the frequency of clinically unsafe outputs by 55.6% compared to conventional semantic similarity retrieval. This substantial reduction in unsafe recommendations marks a significant improvement in the safety risk profile, particularly in settings with limited access to specialized care and clinical guidance, addressing a key safety limitation of generative AI systems in healthcare.

In contrast, semantic similarity-based retrieval exhibited higher variability and unsafe output frequency, highlighting the safety limitations of conventional retrieval strategies. Additionally, automated LLM-as-a-Judge evaluation systematically overestimated safety relative to physician assessment, reinforcing the necessity of human expert validation for safety-critical clinical applications.

These findings suggest that safe clinical document generation using LLMs is achievable when grounded in structured reasoning and clinically constrained retrieval. More broadly, this work shows that retrieval functions act as a safety control layer in generative AI systems, rather than merely a performance optimization.

Importantly, the proposed system is intended as a clinical decision support tool within real clinical workflows rather than a replacement for physician judgment. It supports clinicians and broader populations in low-expertise and resource-constrained settings by providing protocol-grounded guidance while maintaining human oversight in safety-critical scenarios.

Future work should investigate prospective real-world deployment, develop retrieval architectures explicitly optimized for clinical safety, and further expand protocol coverage to additional diseases.

Although evaluated in the context of SCD, these findings generalize to safety-critical applications of generative AI more broadly, where rare but high-impact failures dominate system risk. More broadly, this work suggests that evaluation paradigms for generative AI in healthcare must move beyond average metrics and account for safety-critical failure modes. This shift is essential for responsible and equitable deployment in real-world clinical environments.

Acknowledgments

The authors declare no conflicts of interest. Funded by the **Agents4Good project** (Kunumi[9]; Embrapii CEEI/UFCG Software and Automation Unit[10]).

Approved by the Institutional Research Ethics Committee and registered on *Plataforma Brasil* (CAAE: 89676025.9.0000.5182). Informed consent obtained, with no real patient data used.

[8] https://bit.ly/hemachat

[9] https://www.kunumi.com/br

[10] https://embrapii.org.br/unidades/software-e-automacao-ceei

References

Selwa Y. Abdeldafie and Sameera O. Alaajmi. 2022. Knowledge and attitudes of nurses toward sickle cell disease patients in jazan. *Journal of Family Medicine and Primary Care*, 11(11):6935–6943.

Lameck Mbangula Amugongo, Pietro Mascheroni, Steven Brooks, Stefan Doering, and Jan Seidel. 2025. Retrieval augmented generation for large language models in healthcare: A systematic review. *PLOS Digital Health*, 4(6):1–33.

Muhammad Arslan, Hussam Ghanem, Saba Munawar, and Christophe Cruz. 2024. A survey on rag with llms. *Procedia Computer Science*, 246:3781–3790. 28th International Conference on Knowledge Based and Intelligent information and Engineering Systems (KES 2024).

Pat Croskerry. 2009. A universal model of diagnostic reasoning. *Academic Medicine*, 84(8):1022–1028.

Edward Croxford and 1 others. 2025. Evaluating clinical ai summaries with large language models as judges. *npj Digital Medicine*, 8.

Andrews Adjei Druye, Dorcas Frempomaa Agyare, William Akoto-Buabeng, Jethro Zutah, Frank Odonkor Offei, Bernard Nabe, Godson Obeng Ofori, Amidu Alhassan, Benjamin Kofi Anumel, Godfred Cobbinah, Susanna Aba Abraham, Mustapha Amoadu, and John Elvis Hagan. 2024. Healthcare professionals' knowledge, attitudes, and practices in the assessment, and management of sickle-cell disease: A meta-aggregative review. *Diseases*, 12(7):156.

Jiawei Gu and 1 others. 2024. A survey on llm-as-a-judge. *The Innovation*.

Hospital de Clínicas de Porto Alegre. 2023. Protocolos de emergência pediátrica. UFRGS Institutional Repository.

Hsin-Ling Hsu, Cong-Tinh Dao, Luning Wang, Zitao Shuai, Thao Nguyen Minh Phan, Jun-En Ding, Chun-Chieh Liao, Pengfei Hu, Xiaoxue Han, Chih-Ho Hsu, Dongsheng Luo, Wen-Chih Peng, Feng Liu, Fang-Ming Hung, and Chenwei Wu. 2025. Medplan:a two-stage rag-based system for personalized medical plan generation. *Preprint*, arXiv:2503.17900.

Institute of Medicine. 2011. *Clinical Practice Guidelines We Can Trust*. National Academies Press, Washington, DC.

Gregory J. Kato, Frédéric B. Piel, Clarice D. Reid, Marilyn H. Gaston, Kwaku Ohene-Frempong, Lakshmanan Krishnamurti, Wally R. Smith, Julie A. Panepinto, David J. Weatherall, Fernando F. Costa, and Elliott P. Vichinsky. 2018. Sickle cell disease. *Nature Reviews Disease Primers*, 4(1):18010.

Arshia Kermani, Veronica Perez-Rosas, and Vangelis Metsis. 2025. A systematic evaluation of llm strategies for mental health text analysis: Fine-tuning vs. prompt engineering vs. rag. *Preprint*, arXiv:2503.24307.

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability. Technical report, University of Pennsylvania.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Preprint*, arXiv:2005.11401.

Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, Kai Shu, Lu Cheng, and Huan Liu. 2025. From generation to judgment: Opportunities and challenges of LLM-as-a-judge. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 2757–2791, Suzhou, China. Association for Computational Linguistics.

Fred Mutisya, Shikoh Gitau, Christine Syovata, Diana Oigara, Ibrahim Matende, Muna Aden, Munira Ali, Ryan Nyotu, Diana Marion, Job Nyangena, Nasubo Ongoma, Keith Mbae, Elizabeth Wamicha, Eric Mibuari, Jean Philbert Nsengemana, and Talkmore Chidede. 2025. Mind the gap: Evaluating the representativeness of quantitative medical language reasoning llm benchmarks for african disease burdens. *Preprint*, arXiv:2507.16322.

Fnu Neha, Deepshikha Bhati, and Deepak Kumar Shukla. 2025. Retrieval-augmented generation (rag) in healthcare: A comprehensive review. *AI*, 6(9).

Oluwatosin Ogundare and Subuola Sofolahan. 2023. Large language models in ambulatory devices for home health diagnostics: A case study of sickle cell anemia management. In *Advances in Intelligent Networking and Collaborative Systems*, pages 447–453, Cham. Springer Nature Switzerland.

Frédéric B. Piel, Thomas N. Williams, and David J. Weatherall. 2017. Sickle cell disease. *New England Journal of Medicine*, 376(16):1561–1573.

David C Rees, Thomas N Williams, and Mark T Gladwin. 2010. Sickle-cell disease. *The Lancet*, 376(9757):2018–2031.

Jessie Reich, Mary Ann Cantrell, and Suzanne C. Smeltzer. 2022. An integrative review: The evolution of provider knowledge, attitudes, perceptions and perceived barriers to caring for patients with sickle cell disease 1970–now. *Journal of Pediatric Hematology/Oncology Nursing*, 40(1):43–64.

Sociedade de Pediatria do Estado do Rio de Janeiro. 2023. Protocolos clínicos. Revista SOPERJ.

Song Sun, Zhijie Zhong, Nanlan Yu, Xinrong Gong, and Kaixiang Yang. 2025. Humanmod: A multi-rag collaborative llm for inclusive urban public healthcare services. *Applied Soft Computing*, 184:113684.

Anna Visibelli, Bianca Roncaglia, Ottavia Spiga, and Annalisa Santucci. 2023. The impact of artificial intelligence in the odyssey of rare diseases. *Biomedicines*, 11(3):887.

Dariusz Walkowiak and Jan Domaradzki. 2021. Are rare diseases overlooked by medical education? awareness of rare diseases among physicians in poland: an explanatory study. *Orphanet Journal of Rare Diseases*, 16(1).

C. Wang, S. Liu, H. Yang, J. Guo, Y. Wu, and J. Liu. 2023. Ethical considerations of using chatgpt in health care. *Journal of Medical Internet Research*, 25:e48009.

Guanchu Wang, Junhao Ran, Ruixiang Tang, Chia-Yuan Chang, Chia-Yuan Chang, Yu-Neng Chuang, Zirui Liu, Vladimir Braverman, Zhandong Liu, and Xia Hu. 2024a. Assessing and enhancing large language models in rare disease question-answering. *Preprint*, arXiv:2408.08422.

Leyao Wang, Zhiyu Wan, Congning Ni, Qingyuan Song, Yang Li, Ellen Clayton, Bradley Malin, and Zhijun Yin. 2024b. Applications and concerns of chatgpt and other conversational large language models in health care: Systematic review. *Journal of Medical Internet Research*, 26:e22769.

Yunfan Yan, Chi Zhang, Donghan Yu, Weizhe Lin, Chenlin Meng, Chenyan Xiong, Zhiyuan Liu, Zheng Zhang, Tat-Seng Chua, and Maosong Sun. 2024. Corrective retrieval augmented generation. *Preprint*, arXiv:2401.15884.

Huizi Yu, Jiayan Zhou, Lingyao Li, Shan Chen, Jack Gallifant, Anye Shi, Jie Sun, Xiang Li, Jingxian He, Wenyue Hua, Mingyu Jin, Guang Chen, Yang Zhou, Zhao Li, Trisha Gupte, Ming-Li Chen, Zahra Azizi, Qi Dou, Bryan P. Yan, and 7 others. 2025. Simulated patient systems powered by large language model-based ai agents offer potential for transforming medical education. *Communications Medicine*, 6(1):27.

Zhihong Zhang, Pallavi Gupta, Jiyoun Song, Maryam Zolnoori, and Maxim Topaz. 2025. From conversation to standardized terminology: An llm-rag approach for automated health problem identification in home healthcare. *Journal of Nursing Scholarship*, 57(6):1003–1011.

A Appendix

A.1 Evaluation Pipeline

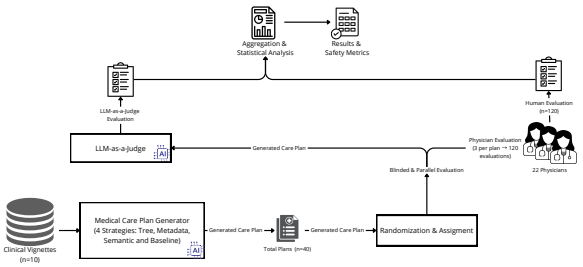


Figure 2: Detailed overview of the experimental evaluation pipeline, including medical care plan generation, randomization, blinded physician assessment, and parallel LLM-as-a-Judge evaluation.

A.2 LLM as a judge prompt evaluator

LLM-as-a-Judge

You are a health specialist evaluator with clinical experience in caring for patients with Sickle Cell Disease.
Your role is to evaluate medical referrals for patients with Sickle Cell Disease.
Your task is to compare a reference medical referral with a medical referral under evaluation,
considering the clinical context of the case, and assign Likert scale scores from 1 to 5 for each evaluation criterion described below.
Likert Scale:
1 – Strongly disagree
2 – Disagree
3 – Neither disagree nor agree
4 – Agree
5 – Strongly agree

INPUTS:
Clinical context of the case (patient with Sickle Cell Disease; symptoms and relevant information):
{CLINICAL CONTEXT}
Reference medical referral:
{ORIGINAL MEDICAL PLAN}
Medical referral to be evaluated:
{GENERATED MEDICAL PLAN}

EVALUATION CRITERIA
- Clinical adequacy: The generated referral is clinically -appropriate for a patient with Sickle Cell Disease, correctly considering the clinical context, symptoms, and described information. The tests, management decisions, and diagnostic hypotheses included in the generated referral are appropriate for managing patients with Sickle Cell Disease and for the presented clinical condition.
- Completeness: The generated referral is complete and includes all relevant management decisions, tests, and clinical observations expected for managing a patient with Sickle Cell Disease, according to the presented case.
- Clinical safety: The generated referral is clinically safe and does not expose the patient with Sickle Cell Disease to risks, considering the diagnostic hypotheses, tests, and management decisions described.

OUTPUT FORMAT (MANDATORY):
The output must be exclusively in JSON format, with no additional text outside the JSON.
{
"adequacao_clinica": <integer from 1 to 5>,
"completude": <integer from 1 to 5>,
"seguranca_clinica": <integer from 1 to 5>
}

IMPORTANT RULES
- Scores must be integers between 1 and 5.
- Evaluate exclusively the medical referral under evaluation.
- Base the evaluation only on the provided clinical context and referrals.
- Do not invent information that is not present in the input data.

Figure 3: Prompt used for the LLM-as-a-Judge evaluation.

A.3 Clinical Evaluation Criteria for Medical Care Plans

- **Clinical Adequacy**, defined as the extent to which the medical care plans appropriately interpreted the clinical scenario and proposed diagnostically coherent and medically appropriate reasoning;
- **Completeness**, defined as whether the medical care plans included all essential diagnostic investigations and therapeutic actions required for safe and effective patient management;
- **Clinical Safety**, defined as the absence of recommendations that could expose the patient to preventable harm, including omissions of critical interventions, inappropriate therapeutic sequencing, or potentially iatrogenic actions.

A.4 Qualitative Physician Feedback on Medical Care Plans

Qualitative feedback from physicians reveals distinct failure modes across retrieval strategies, providing insight beyond average performance.

Baseline outputs frequently exhibited incomplete reasoning, including missing diagnostic steps, insufficient justification of hypotheses, and lack of prioritization of critical interventions.

- *"Missing important diagnostic investigations for proper evaluation of the clinical condition."*
- *"Treatment plan incomplete and lacking prioritization of critical interventions."*
- *"Insufficient justification of diagnostic hypotheses given the clinical presentation."*
- *"The plan appears adequate, but lacks clinical depth for safe decision-making."*

Semantic Similarity Retrieval clinically inappropriate hypotheses and unnecessary interventions, often inconsistent with the clinical scenario, representing a higher-risk failure mode.

- *"Diagnostic hypotheses are not supported by the clinical presentation and lead to unnecessary tests."*
- *"Unnecessary investigations such as chest X-ray and antibiotic therapy were suggested."*

- *"Some recommended actions do not impact clinical outcomes and may delay appropriate management."*
- *"Irrelevant hypotheses were introduced while important alternatives were not considered."*

Metadata-Filtered Retrieval improved structure and interpretability, but still produced occasional inconsistencies and unnecessary procedures.

- *"The plan is clear and sufficiently detailed, allowing appropriate clinical action."*
- *"Some diagnostic hypotheses are not fully supported by clinical findings."*
- *"Unnecessary diagnostic tests may delay clinical decision-making."*

LLM-Guided Tree Retrieval produced more coherent and clinically aligned reasoning. Observed issues were limited to minor refinements, such as occasional unnecessary tests, rather than structural errors.

- *"The clinical reasoning is coherent and aligned with the presented case."*
- *"Minor adjustments could improve the selection of diagnostic tests."*
- *"The management plan is consistent with the clinical scenario."*

Overall, these findings show a clear shift in error type: from structural and clinically inconsistent failures (Baseline and Semantic Retrieval) to refinement-level issues (LLM-Guided Tree Retrieval).

A.5 Distribution table of self-reported evaluator experience in SCD

Table 3: Self-reported evaluator experience in SCD (%)

Experience Level	(1) General	(2) Treatment
Level 1	0.0	0.0
Level 2	9.5	9.5
Level 3	52.4	61.9
Level 4	33.3	23.8
Level 5	4.8	4.8

A.6 Vignette example - Translated to english

The following example illustrates the structure of the clinical vignettes used in the study, along with the corresponding gold-standard medical care plan.

Vignette Example

- ID: J.L.S., 15 years old, born and residing in Campina Grande, diagnosed with sickle cell anemia and under follow-up at the hemocenter, accompanied by his mother.
- CHIEF COMPLAINT: Severe body pain for 2 days.
- HISTORY OF PRESENT ILLNESS: The patient reports severe and continuous pain in the upper limbs for approximately 48 hours, with no history of trauma. He states that the pain did not improve after taking dipyrone at home. Denies fever, cough, dyspnea, abdominal pain, or urinary symptoms. Reports reduced fluid intake and recent exposure to cold temperatures. He is on regular use of hydroxyurea and folic acid, with the last painful crisis occurring 6 months ago. Denies previous surgeries and allergies.
- PHYSICAL EXAMINATION: On examination, the patient is conscious, oriented, and anxious due to pain, pale (+/4+), with HR 104 bpm, BP 110/70 mmHg, RR 20 breaths per minute, Temp. 36.8°C, and SpO_2 96\% on room air. Lung auscultation without abnormalities, normal heart sounds, flat and non-tender abdomen, no visceromegaly. Diffuse pain on palpation of the long bones of the upper limbs, without inflammatory signs.

OUTPUT — Gold-standard medical care plan created by the specialist physician:

- DIAGNOSTIC HYPOTHESES: Painful vaso-occlusive crisis; Occult infection (less likely); Early acute chest syndrome (to be ruled out).
- SUGGESTED COMPLEMENTARY TESTS: Complete blood count, reticulocyte count, urea, creatinine, electrolytes, C-reactive protein, chest X-ray (if respiratory symptoms develop), liver function tests.
- SUGGESTED MANAGEMENT:
• Immediate initiation of stepwise analgesia according to pain intensity, including opioids if refractory to dipyrone/NSAIDs.
• Intravenous hydration if necessary, encourage oral intake.
• Oxygen therapy only if $SpO_2 < 95\%$.
• Monitor analgesic response and vital signs.
• Laboratory investigation to rule out complications.
• Do not indicate routine transfusion (consider only in case of significant hemoglobin drop, acute chest syndrome, or other complication).
• Provide guidance on crisis prevention (adequate hydration, avoiding cold exposure, infection control).

Figure 4: Example of a clinical vignette and its corresponding gold-standard medical care plan used in the evaluation.

A.7 Violin distribution plot of criteria scores for LLM and human evaluation

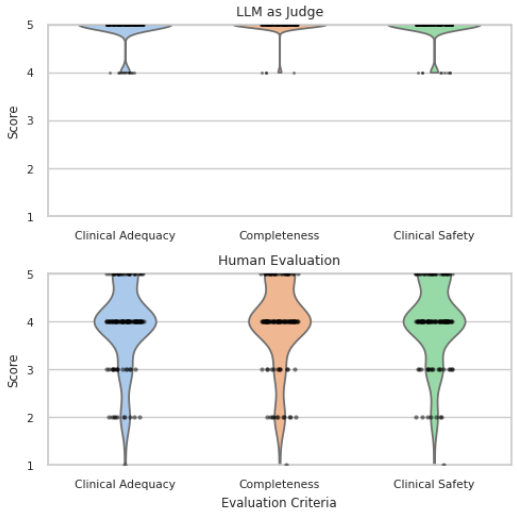


Figure 5: Distribution of criteria scores for LLM and human evaluation

Class of LLMs: Benchmarking Large Language Models on the Brazilian National Medical Examination

João Vitor Mariano Correia[1], **Pedro Henrique Alves de Castro**[2], **Gabriel Lino Garcia**[1], **Pedro Henrique Paiola**[1], **João Paulo Papa**[1]
[1]Department of Computing, Faculty of Sciences, São Paulo State University
[2]Department of Medical Sciences, Nove de Julho University

Abstract

The evaluation of Large Language Models (LLMs) in medicine has predominantly relied on English-language benchmarks aligned with North American clinical guidelines, limiting their applicability to other healthcare systems. In this paper, we evaluate twenty-two proprietary and open-weight LLMs on the 2025 National Examination for the Evaluation of Medical Training (ENAMED), a high-stakes, government-standardized assessment used to evaluate medical graduates in Brazil. The benchmark comprises 90 multiple-choice questions grounded in Brazilian public health policy, clinical practice, and Portuguese medical terminology, and is released as an open dataset. Model performance is measured using both standard accuracy and the official Item Response Theory (IRT) framework employed by ENAMED, enabling direct comparison with human proficiency thresholds. Results reveal a clear stratification of model capabilities: proprietary frontier models achieve the highest performance, whereas many open-weight and smaller-domain-adapted models fail to meet the minimum proficiency criterion. Across comparable scales, large generalist models consistently outperform specialized medical fine-tunes, suggesting that general reasoning capacity is a stronger predictor of success than narrow domain adaptation in this setting. These findings establish ENAMED as a rigorous benchmark for evaluating medical LLMs in Portuguese and highlight both the potential and current limitations of such models for educational assessment.

1 Introduction

The integration of Artificial Intelligence into clinical practice has advanced Large Language Models (LLMs) from experimental systems to evaluated tools for tasks like summarization and question answering. While frontier models now pass global, English-language benchmarks such as the USMLE (Nori et al., 2023; Singhal et al., 2023), these assessments ignore the organizational structure of the **Brazilian Unified Health System** (SUS), regional epidemiology, and the Federal Council of Medicine (CFM) professional norms.

In 2025, Brazil introduced the **National Examination for the Evaluation of Medical Training (ENAMED)**, a **centralized** assessment consolidating undergraduate and residency evaluations. Aligned with National Curricular Guidelines (DCNs), ENAMED uniquely prioritizes **public health policies** and primary care. Its inaugural administration yielded **unsatisfactory** institutional scores, providing a rigorous, government-standardized context for evaluating AI preparedness and clinical reasoning in Brazilian medical education.

This work evaluates the performance of general-purpose and domain-adapted Large Language Models on ENAMED. Our results show that recent generalist models consistently outperform several specialized medical models. The main contributions of this study are twofold: (i) a systematic, domain-level analysis of LLM behavior in a national medical assessment setting, highlighting both their potential and the challenges of deploying such models in context-specific healthcare environments, and (ii) the release of a structured dataset derived from ENAMED 2025, enabling reproducible evaluation and future research on LLM performance in Brazilian medical education.

2 Related Work

Early evaluations of medical LLMs relied primarily on English-language benchmarks, including general-purpose and biomedical question-answering datasets such as MMLU (Hendrycks et al., 2021), PubMedQA (Jin et al., 2019), and MedQA (Jin et al., 2021). Although these datasets effectively assess biomedical knowledge and multi-

Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)
Vol. 2, pages 101–111
April 13-16, 2026

step reasoning, they reflect predominantly English-language, high-resource clinical environments. This limitation has motivated the development of Portuguese-language medical resources. SemClinBR (Oliveira et al., 2022) and BRATECA (Dias and Ulbrich, 2022) established foundational corpora for clinical NLP in Brazilian Portuguese, supporting domain adaptation efforts. Subsequent studies developed and evaluated Portuguese-adapted medical LLMs, demonstrating the feasibility of regional specialization (de Souza Pinto et al., 2024; Paiola et al., 2024).

More recently, benchmark construction has shifted toward examination-based evaluation. HealthQA-BR and related initiatives (D'addario, 2025; Garcia et al., 2025) introduced large-scale benchmarks derived from Brazilian national licensing and residency examinations, revealing substantial variability in model performance across specialties and professional domains. These works underscore that high aggregate accuracy may mask domain-specific weaknesses, particularly in locally grounded regulatory and administrative knowledge.

Our study extends this line of research by introducing a structured version of the 2025 ENAMED examination and systematically benchmarking a broad spectrum of proprietary and open-weight LLMs on this high-stakes assessment. Unlike prior evaluations that report aggregate accuracy alone, we contextualize model performance using the official Item Response Theory framework employed for human examinees, enabling direct comparability with institutional proficiency thresholds. By combining structured dataset release, large-scale comparative benchmarking, and psychometric alignment, this work provides an updated assessment of the current stage of LLM capability in a national, high-stakes medical evaluation setting.

3 Methodology

To evaluate model performance, we constructed a structured dataset based on the 2025 administration of the ENAMED examination. The dataset construction process comprised three phases: data acquisition, automated extraction with human review, and multimodal adaptation[1].

Data Acquisition and Extraction Source materials were obtained from the official INEP repository,[2], consisting of the examination booklet and official answer key in PDF format. A rule-based parsing pipeline converted the unstructured PDFs into machine-readable form by isolating question statements and alternatives, removing document artifacts such as headers and page numbers, and aligning correct answers with the official key.

Data Curation and Cleaning A human-in-the-loop review was conducted to correct residual extraction errors. Answer alternatives fragmented during parsing were manually realigned, clinical tables were converted to Markdown to improve tokenizer compatibility, and questions annulled by the examination board or excluded from official scoring were removed to ensure consistency with evaluation criteria.

Reference Matrix Classification For domain-level analysis, each question was mapped to seven competency areas from the Common Reference Matrix (Ministério da Educação (MEC) and INEP, 2025). Lacking official labels, we employed a model-as-judge consensus (majority vote from `Gemini 3 Pro`, `GPT-5`, and `Sabiá 4`). This yielded almost perfect agreement (Fleiss' $\kappa = 0.82$ (Landis and Koch, 1977)), with the highest pairwise concordance between GPT-5 and Gemini 3 Pro ($\kappa = 0.907$), followed by GPT-5 and Sabiá 4 ($\kappa = 0.798$).

Multimodal Processing A subset of questions ($n = 3$) required interpretation of clinical images. To enable evaluation with text-only models, textual descriptions of the visual stimuli were generated using the `gemini-3-pro-preview` model. These descriptions were reviewed by a medical student to improve clinical clarity, primarily correcting photographic artifacts, anatomical imprecision, and misleading surface attributes. Importantly, this intermediate review step does not constitute a validation of the generated descriptions; the absence of assessment by board-certified clinicians remains a limitation outside the scope of this study.

Final Dataset Statistics After cleaning and filtering, the final dataset comprises 90 multiple-choice questions. Each item is represented in JSON format and includes the question text, four answer options,

[1]The dataset is publicly available at: https://huggingface.co/datasets/recogna-nlp/enamed-2025.

[2]https://www.gov.br/inep/pt-br/areas-de-atuacao/avaliacao-e-exames-educacionais/enamed/provas-e-gabaritos

the ground-truth label, and, where applicable, a generated image description.

4 Experimental Setup

4.1 Computational Infrastructure

All local inference experiments were conducted on a computing node equipped with a single NVIDIA H100 Tensor Core GPU. This hardware configuration supports high-throughput inference for open-weight models through large memory bandwidth and Tensor Core acceleration, being executed using the `HuggingFace Transformers` library. Proprietary models were accessed via their respective official APIs. All models were evaluated in a zero-shot setting using the standardized prompt template detailed in Appendix A.

4.2 Model Selection

To evaluate a range of current Large Language Models in medical reasoning, a cohort of 22 models was selected. The models were stratified into three categories based on access modality and adaptation strategy.

Proprietary Frontier Models These models represent the state of the art (SOTA) in reasoning and knowledge retrieval, trained on massive multilingual corpora and accessible via an API. The **Gemini 3** family (Google) was evaluated, including `gemini-3-pro` and `gemini-3-flash`, both based on mixture-of-experts architectures with multimodal capabilities. The **GPT-5** family (OpenAI) was evaluated, including `gpt-5` and its smaller variants, `gpt-5-mini` and `gpt-5-nano`, to examine the relationship between model scale and performance. We also included proprietary models pre-trained specifically on Portuguese data, rather than merely fine-tuned. We selected the **Sabiá 4 Family** (Maritaca AI), a suite of models specialized for the Lusophone context. We evaluated `Sabia 4`, designed for complex reasoning tasks, and `Sabiazinho 4`, optimized for low-latency applications. These models provide a reference point for evaluating the impact of Portuguese-language specialization relative to large general-purpose models.

Open-Weight Generalist Models We selected a range of state-of-the-art open-source models to evaluate the capabilities of accessible AI on local hardware without any specific medical adaptation. This cohort includes the **Qwen 3** family (Yang et al., 2025), known for strong multilingual performance, where we tested the 14B, 8B, and 4B parameter variants to observe scaling laws. We also included **Phi-4** (Abdin et al., 2024), a compact reasoning model (15B and 4B) trained heavily on high-quality synthetic data, and **Llama 3.1** (Grattafiori et al., 2024), specifically the 8B parameter variant, which serves as the standard baseline for general-purpose open LLMs.

Open-Weight Adapted Models This category encompasses open models that have undergone post-training adaptation for specific domains to test the efficacy of fine-tuning versus pre-training. In the medical domain, we evaluated **MedGemma** (Sellergren et al., 2025), built upon the Gemma 3 architecture (27B and 4B) and fine-tuned on biomedical corpora such as PubMed and MIMIC-III, as well as `MMed-Llama-3` (Qiu et al., 2024), an 8B model further pre-trained on the MMedC multilingual medical corpus. Regarding Portuguese and clinical adaptations, we assessed the **Bode Family** (Paiola et al., 2025), including `Bode 3.1-8B` (general Portuguese adaptation) and `DrBode-240k` (medical fine-tuning on Brazilian clinical cases). Finally, to benchmark progress over previous generations, we included legacy baselines such as `Sabiá-7B` (Pires et al., 2023)(based on Llama 1) and `Clinical-BR-LlaMA-2-7B` (de Souza Pinto et al., 2024).

4.3 Evaluating Measures

Model performance was evaluated using both standard classification metrics and the official psychometric framework adopted by ENAMED. The latter is based on Item Response Theory (IRT) and enables estimation of model proficiency (θ) on the same latent scale used to assess human examinees.

Standard Classification Metrics We model the task as a four-way multiple-choice classification problem, where each model predicts an answer $\hat{y}_i \in \{A, B, C, D\}$ for item i with ground-truth label y_i. We report Accuracy and Macro-F1, defined respectively as $\frac{1}{N}\sum_i \mathbb{I}(\hat{y}_i = y_i)$ and the unweighted mean of per-class F1 scores over $\{A, B, C, D\}$. Divergences between Accuracy and Macro-F1 are interpreted as indicators of asymmetric class behavior.

Psychometric Evaluation Beyond standard classification metrics, performance was evaluated using the official ENAMED psychometric framework based on the Rasch (1PL) Item Response Theory

model, enabling direct comparison with human examinees. In the Rasch model, each item i is characterized by a difficulty parameter b_i, defined as the proficiency level at which the probability of a correct response is 50%. For a model j with latent proficiency θ_j, the probability of correctly answering item i is:

$$P(U_{ij} = 1 \mid \theta_j, b_i) = \frac{1}{1 + e^{(b_i - \theta_j)}}. \quad (1)$$

Both θ_j and b_i are defined on the same logit scale. We used the official item difficulty parameters provided by INEP, without recalibration, consistent with the examination's 1PL scaling framework.

Model proficiency was estimated using the IRT True-Score (TS) estimator adopted by INEP. For each model, the observed raw score was mapped to the corresponding latent proficiency value by inverting the test characteristic curve implied by the Rasch model parameters. Estimation was conducted over the standard interval $[-4, 4]$, consistent with official technical documentation. The latent scale ($\mu = 0$, $\sigma = 1$) was then linearly transformed to the ENAMED reporting scale. Following the Modified Angoff procedure, the minimum proficiency threshold corresponds to $\theta = -0.40$, equivalent to a score of 60.0.

Institutional Concept (Enade Concept) We evaluate the models collectively using the **Enade Concept**, the 1–5 categorical rating employed by the Brazilian Ministry of Education to assess medical schools. The concept is determined by the proportion of evaluated models meeting the official proficiency threshold, which is mapped to discrete levels: Level 1 ($< 40\%$), Level 2 (40%-59%), Level 3 (60%-74%), Level 4 (75%-89%), and Level 5 ($\geq$ 90%). This metric summarizes the proportion of evaluated models meeting the proficiency threshold relative to institutional benchmarks used in medical education.

5 Results

Table 1 presents the evaluation metrics on the ENAMED dataset. The models are ranked by global accuracy, revealing distinct performance tiers driven by model scale, architecture, and training methodology.

The results reveal a clear stratification of capabilities across model classes. Proprietary frontier architectures, particularly the **Gemini-3** and **GPT-5** families, consistently achieved the highest performance on the medical benchmark, forming a distinct upper tier. `Gemini-3-pro` attained the best overall accuracy (98.89%), followed closely by `GPT-5` and `Gemini-3-flash`, all of which exhibited near-ceiling performance. Notably, the Portuguese-centric proprietary model `Sabiá 4` achieved 93.33% accuracy, surpassing `GPT-5-mini` and competing with other high-performing proprietary variants. Accuracy and Macro-F1 were nearly identical across models, indicating minimal class imbalance effects.

A second performance tier comprises primarily efficiency-oriented proprietary variants and regionally optimized models. `GPT-5-mini` (91.11%) and the Brazilian-specialized `Sabiazinho 4` (87.78%) demonstrated strong performance, with `Sabiazinho 4` marginally outperforming `GPT-5-nano` (86.67%). While these models also outperformed the evaluated open-weight generalist baselines, such as `Qwen3-14b` (81.11%), this comparison should be understood as conditional on the model scales examined in this study. Specifically, the open-weight results reflect mid-scale configurations rather than the largest available variants. Accordingly, the observed dominance of proprietary models is robust within the evaluated regime, but should not be interpreted as a definitive comparison against the full upper bound of open-weight architectures.

5.1 Impact of Model Scaling

The evaluation reveals consistent performance stratification within individual model families, indicating that reductions in model capacity or deployment class are associated with measurable losses in medical reasoning performance. Importantly, these observations reflect intra-family trends under the configurations evaluated, rather than a universal comparison across all possible model scales.

Within the proprietary **GPT-5** family, a clear tiering is observed across efficiency-oriented variants. Performance decreases from 97.78% for `GPT-5` to 91.11% for `GPT-5-mini` and further to 86.67% for `GPT-5-nano`. Although architectural details and parameter counts are not publicly disclosed, this pattern suggests systematic trade-offs between deployment efficiency and reasoning capability within a single proprietary model lineage.

In contrast, open-weight families exhibit sharper performance degradation as model size decreases. For the **Qwen3** family, accuracy declines substan-

Model	Raw Score	Accuracy	Macro-F1	ENAMED Score	Proficiency
Gemini 3 Pro Preview	**89**	**0.9889**	**0.9886**	**137.19**	Proficient
Gemini 3 Flash Preview	88	0.9778	0.9773	131.87	Proficient
GPT-5	88	0.9778	0.9778	131.87	Proficient
Sabia 4	84	0.9333	0.9325	110.34	Proficient
GPT-5-mini	82	0.9111	0.9109	104.24	Proficient
Sabiazinho 4	79	0.8778	0.8779	97.13	Proficient
GPT-5-nano	78	0.8667	0.8676	95.11	Proficient
Qwen 3 14B	73	0.8111	0.8123	86.54	Proficient
MedGemma 27B	69	0.7667	0.7677	80.87	Proficient
Phi 4 15B	68	0.7556	0.7527	79.56	Proficient
Gemma 3 12B	62	0.6889	0.6873	72.31	Proficient
LLaMA 3.1 8B	60	0.6667	0.6660	70.07	Proficient
Qwen 3 8B	59	0.6556	0.6520	68.97	Proficient
Bode 3.1 8B	56	0.6222	0.6212	65.74	Proficient
MedGemma 4B	54	0.6000	0.5990	63.64	Proficient
Qwen 3 4B	53	0.5889	0.5892	62.60	Proficient
MMed-Llama-3-8B	46	0.5111	0.4653	55.43	Not Proficient
Gemma 3 4B	43	0.4778	0.4648	52.37	Not Proficient
DrBode 240k	42	0.4667	0.4616	51.35	Not Proficient
Phi 4 Mini 4B	40	0.4444	0.4230	49.29	Not Proficient
Sabiá-7B	32	0.3556	0.3455	40.75	Not Proficient
Clinical-BR-LlaMA-2-7B	Failed to perform				

Table 1: Performance of Large Language Models on the ENAMED 2025 examination. Models are ranked by the official psychometric score (ENAMED Score).

tially from 81.11% in the 14B model to 65.56% and 58.89% in the 8B and 4B variants, respectively. A similar trend is observed in the **Phi-4** family, where the 15B model achieves 75.56% accuracy, while the 4B variant drops to 44.44%. These results indicate that conventional size-reduction strategies in open-weight models are associated with pronounced losses in diagnostic performance under the evaluated conditions.

We limited the analysis of open-weight models to configurations that could be executed within available GPU resources. Consequently, the observed scaling trends do not reflect the full performance potential of larger open-weight architectures, such as higher-parameter variants of the Qwen or LLaMA families. Within these constraints, proprietary models consistently maintain higher performance across multiple deployment tiers, whereas open-weight models exhibit steeper degradation as scale decreases.

5.2 Generalist vs. Domain-Adapted Models

The results reveal that performance differences between generalist and domain-adapted models are strongly mediated by underlying architecture and pre-training regime, rather than by domain specialization alone. In particular, the `Qwen3` family consistently outperforms both `Gemma` and `Phi` models at comparable or smaller parameter scales, suggesting that architectural design choices and multilingual pre-training confer a substantial advantage in medical reasoning tasks conducted in Portuguese.

This effect is evident in the comparison between `MedGemma-27B` and `Qwen3-14B`. Despite having nearly twice the parameter count and being explicitly fine-tuned for the medical domain, `MedGemma-27B` (76.67%) is outperformed by the generalist `Qwen3-14B` (81.11%). However, domain adaptation is not without benefit: `MedGemma-4B` (60.00%) exhibits a measurable improvement over its base counterpart, `Gemma-3-4B` (47.78%), indicating that medical fine-tuning can partially compensate for architectural and scale limitations, particularly in smaller models.

A contrasting pattern emerges among older, domain-adapted Portuguese and medical models. `Clinical-BR-LLaMA-2-7B`, `Sabiá-7B`, and `DrBode-240K` are all based on pre-2023 architec-

tures, which lack the instruction-following fidelity, multilingual robustness, and reasoning capacity of more recent model families. These architectural constraints appear to outweigh the benefits of language or domain adaptation, leading to substantially degraded performance. In some cases, most critically in `Clinical-BR-LLaMA-2-7B`, these models fail to reliably follow task instructions or produce valid answer selections, rendering them unsuitable for this benchmark.

Overall, these findings indicate that domain or language specialization alone is insufficient to overcome limitations imposed by outdated architectures or weaker pre-training regimes. Modern architectural advances and strong general reasoning capabilities remain prerequisites for effective domain adaptation in medical NLP tasks.

5.3 Institutional Performance Assessment

To provide a holistic assessment of the current state of Generative AI in medicine, we applied the official Enade Concept methodology to our set of evaluated models, treating them as a single "graduating cohort" of medical students. As shown in Table 1, 16 out of the 22 evaluated models achieved the minimum proficiency threshold, with one of them failing to perform and being taken out of the evaluation. According to the official conversion scale, where Level 1 represents $< 40\%$ proficiency and Level 5 represents $\geq 90\%$, our "LLM Class of 2025" falls within the $75\% - 90\%$ range.

Consequently, the aggregate performance of current Large Language Models achieves Enade Concept 4. This corresponds to a high-quality performance level under official criteria. It is critical to note that this equivalence is strictly confined to psychometric performance under the ENAMED evaluation framework and does not imply equivalence in supervised clinical practice, ethical judgment, or patient interaction competencies.

5.4 Error Analysis and Qualitative Evaluation

An item-level analysis reveals that model failures are not uniformly distributed but cluster around specific hard questions.

Alignment with Human Psychometrics To assess whether the difficulty perceived by LLMs aligns with human psychometric benchmarks, we analyzed the correlation between the official Item Difficulty (b) and the aggregated model accuracy (Figure 1). We observed a moderate negative correlation ($\rho = -0.4698$, $p < 0.001$). While the directionality indicates that models generally perform worse on items with higher discrimination parameters, the magnitude of the correlation indicates a pronounced alignment gap. Unlike human candidates, whose performance degrades linearly with item complexity, LLMs exhibit a jagged difficulty curve, often solving hard memorization-based items while failing easy context-dependent questions.

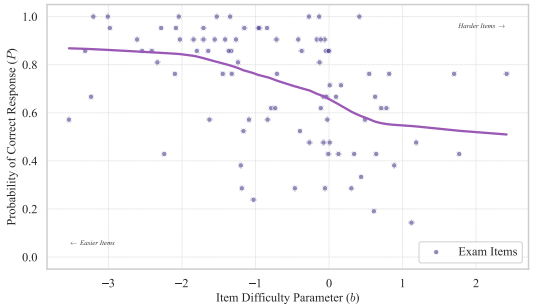


Figure 1: Empirical Item Characteristic Curve comparison. The scatter plot illustrates the probability of correct response for the models as a function of the official item difficulty (b). The trend line indicates a moderate negative correlation ($\rho = -0.4698$), highlighting the divergence between human and machine perceptions of difficulty.

Consensus Errors A significant finding is the presence of consensus errors, where models uniformly hallucinate the same incorrect procedure (e.g., Q61, Q20, Q23). The most prominent example is Question 83, which deals with cervical cancer screening guidelines (see Appendix Table 3 for the full clinical vignette). In this scenario, 100% of the failing models selected "Colposcopy" over the correct protocol of repeating the exam. This error highlights a critical misalignment between general medical knowledge and specific Brazilian public health protocols. While international or private practice guidelines might suggest immediate investigation for LSIL, the *Brazilian Guidelines for Cervical Cancer Screening* (Ministry of Health/SUS) explicitly recommend a conservative approach for LSIL in this age group. The models exhibited a bias for action, preferring an interventional procedure over the correct watchful waiting protocol, likely driven by training data dominated by diverse international guidelines rather than localized SUS protocols.

Multimodal Disparity While most image-based questions (e.g., Q4, Q5, Q53) were solved by nearly all models, Question 96 emerged as a significant outlier. For this item, the text description and the accompanying image clearly describe a *soft, painful* ulcer with irregular borders (characteristic of Chancroid). However, models frequently predicted Syphilis, which typically presents as a *hard, painless* chancre. This suggests a frequency bias in multimodal reasoning: the models likely ignored the fine-grained visual/textual semiotics provided in the vignette and defaulted to the most common statistical cause of genital ulcers (Syphilis). Possibly, current multimodal LLMs may struggle to ground specific visual features when they contradict a strong prior probability heuristic.

The Solved Subset Approximately 15% of the exam (e.g., Q1, Q8, Q77) achieved a difficulty score of 0.0, being answered correctly by every single model. A qualitative review reveals that these questions often pertain to medical ethics, the humanities, or highly standardized trauma protocols. For instance, Question 77 assesses cultural competence in the treatment of Indigenous populations (Tikuna ethnicity). All models correctly identified the need to "recognize the knowledges and practices" of the population. This uniform success likely stems from the extensive Reinforcement Learning from Human Feedback (RLHF) applied to modern LLMs, which heavily penalizes insensitivity and rewards culturally competent, empathetic responses. Similarly, questions such as Q93 (Chemical Eye Burn), which requires immediate irrigation, constitute algorithmic medical knowledge in which the standard of care is universal and unambiguous, thereby minimizing the risk of hallucination.

5.5 Domain-Specific Performance

Performance stratified by medical domain reveals substantial heterogeneity that is obscured by aggregate accuracy. Table 2 reports mean accuracy and dispersion across the seven competency areas defined by the ENAMED reference matrix.

General Surgery and Mental Health exhibit the highest average performance across models, whereas Pediatrics constitutes the most challenging domain, with a mean accuracy below 60%. This pattern is consistent across both proprietary and open-weight architectures, suggesting that pediatric reasoning and presentation may pose systematic challenges for current LLMs rather than model-specific weaknesses.

Medical Specialty	**Accuracy** ($\mu \pm \sigma$)	**Nº Items**
General Surgery	0.791 ± 0.408	11
Mental Health	0.779 ± 0.416	10
Family and Community Medicine	0.755 ± 0.431	13
Medical Clinic	0.752 ± 0.432	27
Collective Health	0.689 ± 0.464	6
Gynecology and Obstetrics	0.653 ± 0.476	16
Pediatrics	0.580 ± 0.494	17

Table 2: Model performance by medical domain on the ENAMED 2025 dataset.

Beyond average accuracy, domain-wise variance exposes marked differences in model reliability. Frontier models such as `GPT-5` and `Gemini 3 Pro`, and `Gemini 3 Flash` display comparatively low dispersion across specialties, indicating stable generalist behavior even in lower-performing domains. In contrast, several open-weight and regionally adapted models exhibit highly uneven performance profiles, achieving near-ceiling accuracy in some domains while failing substantially in others.

This effect is particularly pronounced in Collective Health. Brazilian-trained models, including `Sabiá 4` and `Sabiazinho 4`, consistently outperform larger international models in this domain, reflecting the heavy reliance of ENAMED public health items on SUS-specific legislation and administrative frameworks. Models lacking localized pre-training, such as `Phi-4` and `MedGemma`, show sharp performance degradation despite strong results in clinically oriented domains.

These findings indicate that while clinical reasoning abilities transfer relatively well across languages and health systems, institutional and legal medical knowledge remains highly localized. Consequently, robust performance on national licensing examinations requires either targeted regional pre-training or explicit integration of local knowledge sources, particularly for deployment in public health and policy-sensitive settings.

5.6 Environmental Impact

To ensure that the pursuit of high-performance medical AI aligns with ecological sustainability, we estimated the carbon footprint of our evaluation using the Machine Learning Impact calculator (Lacoste et al., 2019). Experiments were conducted on a single NVIDIA H100 GPU within the Brazilian National Interconnected System, which has a grid emission factor of 0.0461 $kgCO_2/kWh$ due to the

predominance of hydroelectric and other renewable sources.

The total estimated footprint for the full benchmark was ≈ 0.3 kgCO_2eq. This minimal impact demonstrates that high-stakes medical evaluation need not entail high environmental costs. By leveraging Brazil's low-carbon energy matrix, we show that hosting open-weight models locally offers a sustainable alternative to carbon-intensive querying of global API endpoints, demonstrating that medical proficiency can be achieved without compromising environmental stewardship.

6 Conclusion

This study presented a comprehensive evaluation of twenty-two Large Language Models on the inaugural 2025 ENAMED, a high-stakes benchmark unifying undergraduate assessment and residency selection in Brazil. By testing a diverse cohort ranging from proprietary frontier models to open-weight and domain-specific architectures, we assessed the readiness of these systems to interpret complex clinical vignettes, adhere to Portuguese terminological nuances, and navigate the specific public health guidelines of the Unified Health System (SUS).

The results indicate a clear stratification of capabilities, with proprietary models demonstrating near-ceiling performance on the majority of textual items, although residual systematic errors remain. The proprietary frontier models, specifically Gemini-3-pro, GPT-5, and Sabia 4, achieved accuracies of 98.89%, 97.78%, and 93.33%, respectively, approaching the theoretical ceiling of the examination. This performance far exceeds the threshold of unsatisfactory performance noted in a significant portion of medical universities. These results position LLMs primarily as tools for benchmarking and controlled educational support, rather than for autonomous clinical decision-making. Despite high aggregate performance, the presence of systematic and consensus errors, coupled with domain-specific variability, precludes reliable unsupervised use in real-world medical settings. Furthermore, current LLMs lack clear accountability mechanisms, reinforcing that their outputs must remain under human supervision, particularly in high-stakes clinical contexts.

Critically, our findings challenge the prevailing assumption that domain-specific fine-tuning is strictly necessary for medical proficiency. Generalist models showed capable of outperforming medically adapted counterparts when parameter count was held constant. This suggests that for standardized examinations, the reasoning capabilities derived from pre-training outweigh the benefits of targeted biomedical adaptation on smaller architectures. Additionally, the environmental analysis highlights a distinct advantage for the Brazilian ecosystem: while proprietary APIs offer peak accuracy, local inference of open-source models benefits from Brazil's low-carbon energy grid, providing a sustainable path to national technological independence.

However, the transition from examination passing to clinical utility is not without risks. The identification of consensus errors, in which distinct model families confidentially generated the same incorrect distractor, suggests latent biases or shared misconceptions in pre-training corpora that require auditing. Finally, performance heterogeneity across domains (Table 2) indicates that LLMs should not be assumed to generalize uniformly across all areas of medicine, particularly in policy and region-specific contexts.

Limitations

The primary limitation of this study lies in the multimodal evaluation methodology. Because there is no board-certified medical validation of AI-generated descriptions of clinical images, performance on visual questions serves as a proxy rather than a definitive assessment of visual diagnostic capability. Future work should prioritize end-to-end multimodal evaluation using raw image inputs and expand the error analysis to qualitative audits of "consensus failures" to ensure these systems are robust enough for real-world deployment in the Brazilian healthcare context

We also note the relatively small sample size ($N = 90$), given that ENAMED 2025 is the inaugural edition of the examination. While this limited volume restricts the granularity of subgroup analyses and results in wider statistical confidence intervals, it simultaneously ensures a high degree of data sanity. Future work should prioritize expanding this dataset as subsequent editions of ENAMED are released, allowing for longitudinal tracking of LLM performance. Additionally, the development of end-to-end multimodal evaluation pipelines is necessary to rigorously assess visual diagnostic capabilities without reliance on text-based proxies.

References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J. Hewett, Mojan Javaheripi, Piero Kauffmann, James R. Lee, Yin Tat Lee, Yuanzhi Li, Weishung Liu, Caio C. T. Mendes, Anh Nguyen, Eric Price, Gustavo de Rosa, Olli Saarikivi, and 8 others. 2024. Phi-4 technical report. *Preprint*, arXiv:2412.08905.

Andrew Maranhão Ventura D'addario. 2025. Healthqa-br: A system-wide benchmark reveals critical knowledge gaps in large language models. *Preprint*, arXiv:2506.21578.

João Gabriel de Souza Pinto, Andrey Rodrigues de Freitas, Anderson Carlos Gomes Martins, Caroline Midori Rozza Sawazaki, Caroline Vidal, and Lucas Emanuel Silva e Oliveira. 2024. Developing resource-efficient clinical llms for brazilian portuguese. In *Proceedings of the 34th Brazilian Conference on Intelligent Systems (BRACIS)*. In press.

Henrique Dias and Ana Helena Dias Pereira dos Ulbrich. 2022. BRATECA (Brazilian Tertiary Care Dataset): a Clinical Information Dataset for the Portuguese Language. *PhysioNet*. Version 1.1.

Gabriel Lino Garcia, Joao Renato Ribeiro Manesco, Pedro Henrique Paiola, Pedro Henrique Crespan Ribeiro, Ana Lara Alves Garcia, and Joao Paulo Papa. 2025. A Step Forward for Medical LLMs in Brazilian Portuguese: Establishing a Benchmark and a Strong Baseline . In *2025 IEEE 38th International Symposium on Computer-Based Medical Systems (CBMS)*, pages 214–219, Los Alamitos, CA, USA. IEEE Computer Society.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14).

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. PubMedQA: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2567–2577, Hong Kong, China. Association for Computational Linguistics.

Alexandre Lacoste, Alexandra Luccioni, Victor Schmidt, and Thomas Dandres. 2019. Quantifying the carbon emissions of machine learning. *arXiv preprint arXiv:1910.09700*.

J. R. Landis and G. G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174. PMID: 843571.

Ministério da Educação (MEC) and INEP. 2025. Portaria n° 478, de 18 de julho de 2025: Dispõe sobre a implementação da matriz de referência comum para a avaliação da formação médica. Diário Oficial da União. Accessed: 2026-02-07.

Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. 2023. Capabilities of gpt-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*.

Lucas Emanuel Silva e Oliveira, Ana Carolina Peters, Adalniza Moura Pucca Da Silva, Caroline Pilatti Gebeluca, Yohan Bonescki Gumiel, Lilian Mie Mukai Cintho, Deborah Ribeiro Carvalho, Sadid Al Hasan, and Claudia Maria Cabral Moro. 2022. Semclinbr-a multi-institutional and multi-specialty semantically annotated corpus for portuguese clinical nlp tasks. *Journal of Biomedical Semantics*, 13(1):13.

Pedro Henrique Paiola, Gabriel Lino Garcia, João Vitor Mariano Correia, João Renato Ribeiro Manesco, Ana Lara Alves Garcia, and João Paulo Papa. 2025. The bode family of large language models: Investigating the frontiers of llms in brazilian portuguese. *Journal of the Brazilian Computer Society*, 31(1):917–938.

Pedro Henrique Paiola, Gabriel Lino Garcia, João Renato Ribeiro Manesco, Mateus Roder, Douglas Rodrigues, and João Paulo Papa. 2024. Adapting llms for the medical domain in portuguese: A study on fine-tuning and model evaluation. *Preprint*, arXiv:2410.00163.

Ramon Pires, Hugo Abonizio, Thales Sales Almeida, and Rodrigo Nogueira. 2023. Sabiá: Portuguese large language models. In *Intelligent Systems*, pages 226–240, Cham. Springer Nature Switzerland.

Pengcheng Qiu, Chaoyi Wu, Xiaoman Zhang, Weixiong Lin, Haicheng Wang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2024. Towards building multilingual language model for medicine. *Preprint*, arXiv:2402.13963.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, and 1 others. 2025. Medgemma technical report. *arXiv preprint arXiv:2507.05201*.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, Mike Schaekermann, Amy Wang, Mohamed Amin, Sami Lachgar, Philip Mansfield, Sushant Prakash, Bradley Green, Ewa Dominowska, Blaise Aguera y Arcas, and 12 others. 2023. Towards expert-level medical question answering with large language models. *Preprint*, arXiv:2305.09617.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

A Evaluation Prompt

To enhance reproducibility, we provide the Python function used to generate the zero-shot prompts. The prompt was designed in Portuguese to match the examination language and enforces a strict JSON output format to facilitate automated parsing.

```
Você é um médico especialista prestando
o Exame Nacional de Avaliação da
Formação Médica. Leia a questão
abaixo cuidadosamente e identifique a
alternativa correta.

QUESTÃO: { enunciado }

{% if img_descricao %}
DESCRIÇÃO DA IMAGEM: { img_descricao }
{% endif %}

ALTERNATIVAS:
(A) alternativas['A']
(B) alternativas['B']
(C) alternativas['C']
(D) alternativas['D']

INSTRUÇÕES:
1. Analise o caso clínico e a descrição
da imagem (se houver).
2. Responda APENAS com o formato JSON
contendo a letra da alternativa correta.
3. Não forneça explicações, apenas a
resposta.

Formato de Resposta: {"answer": "A"}
```

Figure 2: The zero-shot prompt template used for the evaluation, presented in Jinja2 syntax. The light purple background distinguishes the input prompt from the academic prose. The prompts are in Portuguese to align with the examination language.

B Qualitative Analysis of Representative Items

The following table presents a subset of the analyzed items, categorized by the error phenomenology observed in the Large Language Models. The "Model Consensus" column indicates the incorrect option most frequently selected by the models (Distractor), while "Ground Truth" indicates the official correct answer.

ID	Clinical Vignette	Model Prediction	Correct
	Consensus Errors (Systematic Misconceptions)		
83	Mulher de 32 anos, sexualmente ativa, comparece à consulta com o médico de família e comunidade para realização do seu primeiro exame preventivo. O médico realiza a coleta de citologia oncótica. Após 3 semanas, a paciente retorna com o resultado "presença de lesão intraepitelial de baixo grau". Considerando esse resultado, qual é a conduta adequada do médico?	Encaminhar para a realização de colposcopia	Repetir o exame citopatológico em 6 meses
61	Mulher de 35 anos, diabética, com laqueadura tubária bilateral, procurou atendimento médico com queixa de prurido genital e disúria terminal, com 7 dias de evolução. Recentemente, fez uso de antibiótico para tratamento de abscesso dental. Ao exame especular, notava-se edema vulvar, hiperemia, fissura, corrimento esbranquiçado e teste das aminas negativo. Com base no agente etiológico mais provável, o tratamento é	metronidazol, 1 aplicador, via vaginal, por 10 noites.	miconazol, 1 aplicador, via vaginal, por 7 noites.
20	Mulher travesti de 28 anos, profissional do sexo, comparece à Unidade Básica de Saúde (UBS) em demanda espontânea. Relata relações sexuais frequentes com diferentes parceiros, com uso inconsistente de preservativos, principalmente durante relações anais receptivas. Há 2 dias teve uma relação sexual desprotegida com um cliente que se recusou a usar camisinha. Nunca utilizou medicamento para profilaxia pré-exposição (PrEP) ou pós-exposição (PEP) à infecção pelo HIV. Considerando que a paciente está assintomática no momento, qual a melhor estratégia de prevenção?	Oferecer teste rápido para HIV e sífilis; prescrever PrEP de início imediato; orientar sobre as vacinas disponíveis no SUS para seu grupo populacional.	Realizar testagem rápida para HIV e sífilis; prescrever PEP mediante resultado não reagente para HIV e programar início da PrEP após término da PEP.
	Ambiguity & Entropy		
96	Homem de 30 anos chega para consulta em Unidade Básica de Saúde (UBS) devido à astenia e úlcera no pênis. Trabalha como profissional do sexo e nem sempre faz uso de preservativo. Há cerca de 3 meses, vem notando emagrecimento (10 kg no período), astenia, febre baixa sem horário fixo e, há 1 semana, observou o aparecimento de úlcera dolorosa no pênis. Nega secreção uretral. Ao exame físico, apresenta-se emagrecido, com uma lesão ulcerada com bordas elevadas sem secreção de aproximadamente 3 centímetros logo abaixo da glande, rasa e de base mole, além de linfonodomegalia inguinal direita, com sinais inflamatórios, sem fistulização	Veneral Disease Research Laboratory (VDRL); reagente; benzilpenicilina benzatina 1,2 milhão de unidades, intramuscular, dose única.	microscopia de esfregaço do fundo da úlcera; Gram negativos agrupados em correntes; azitromicina 500 mg, via oral, 2 comprimidos em dose única.
17	Paciente de 20 anos, sexo masculino, vítima de colisão "automóvel a muro", sem cinto de segurança, é atendido ainda na cena pelo Serviço Móvel de Atendimento de Urgência (SAMU). Exame físico: paciente torporoso; saturação de O2 de 60%, em ar ambiente; frequência respiratória de 28 irpm; frequência cardíaca de 112 bpm; pressão arterial de 90 x 50 mmHg. Desvio da traqueia para a direita, turgência de veias jugulares, hipofonese de bulhas cardíacas e diminuição acentuada do murmúrio vesicular à esquerda. Qual é a conduta adequada no atendimento pré-hospitalar?	Reposição volêmica.	Toracocentese
	Solved (Ceiling Effect)		
77	Uma equipe de saúde da família realiza atendimento itinerante a comunidades ribeirinhas e aldeias indígenas na Região Amazônica. Em visita, uma médica recém-chegada observa que uma mulher ribeirinha evita contato visual durante a consulta e responde às perguntas apenas com monossílabos. Em outra situação, um indígena da etnia Tikuna não aceita ser atendido sozinho e insiste na presença de um pajé da comunidade.	promover espaços formativos para a equipe assistencial, reconhecendo saberes e práticas das populações atendidas.	
93	Paciente de 45 anos atendida na Unidade Básica de Saúde (UBS) com dor ocular. Referiu que estava realizando limpeza doméstica com alvejante e deixou atingir o olho, acidentalmente. Ao exame físico, foi observada presença de hiperemia intensa com opacidade da córnea e queimadura química da pálpebra superior do olho direito. Qual é o correto manejo da paciente?	Lavagem ocular com solução fisiológica e avaliação imediata do especialista.	

Table 3: Qualitative analysis of representative items. Columns 3 and 4 show the divergence between the consensus of the models (Distractor) and the official ground truth.

Retrieval-Augmented Generation for Clinical Question Answering in Portuguese Drug Leaflets: Benefits and Limitations

Gabriel Lino Garcia[1], Pedro Henrique Paiola[1], João Vitor Mariano Correia[1], Douglas Rodrigues[1], João Paulo Papa[1],

[1] São Paulo State University (UNESP)
Av. Eng. Luís Edmundo Carrijo Coube, 14-01 - Bauru - SP - Brazil ,
Correspondence: gabriel.lino@unesp.br

Abstract

Retrieval-Augmented Generation (RAG) is proposed to reduce hallucination and improve grounding in clinical language models, yet its effectiveness across different levels of clinical reasoning remains unclear. We conducted a controlled evaluation of medication-related question answering in Portuguese using over 7,000 Brazilian regulatory drug leaflets and a complementary clinical benchmark derived from national medical licensing examinations (Revalida and Fuvest). Retrieval substantially improved factual recall and clinical coherence in medication-specific queries, increasing F1 from 0.276 to 0.412. However, naive retrieval did not consistently improve complex clinical reasoning and sometimes reduced accuracy compared to a parametric-only baseline. We identify retrieval-induced anchoring bias, where partially relevant evidence shifts model decisions toward clinically incorrect conclusions. Critique-based and adaptive retrieval mitigated this effect and achieved the highest clinical benchmark accuracy (54.25%). Clinically grounded evaluation dimensions revealed safety-relevant differences beyond traditional NLP metrics. These results show that retrieval augmentation is effective in regulatory settings but requires adaptive control for higher-level clinical reasoning.

1 Introduction

Access to reliable and up-to-date medication information remains a persistent challenge in clinical practice, particularly in low- and middle-income countries such as Brazil. Drug leaflets issued by regulatory agencies constitute the official, legally binding source of information on indications, contraindications, dosage, pharmacodynamics, and adverse effects. However, their length, technical density, and fragmented organization frequently hinder efficient information retrieval by healthcare professionals and patients. Difficulties navigating regulatory documentation may increase reliance on secondary summaries or informal sources, which are not always complete, up to date, or clinically safe.

In recent years, large language models (LLMs) have demonstrated remarkable performance in medical question answering, clinical reasoning, and decision support tasks (Singhal et al., 2023). Despite these advances, LLMs operate primarily through parametric knowledge acquired during large-scale pretraining (Brown et al., 2020). When deployed without explicit grounding mechanisms, they are prone to hallucinations, outdated recommendations, and unverifiable justifications (Ji et al., 2023). These limitations are particularly concerning in high-stakes medical environments, where incorrect or poorly supported outputs may lead to harmful clinical decisions. Empirical studies have shown that fluent responses generated by LLMs can contain subtle but clinically significant inaccuracies, reinforcing concerns regarding their safe adoption in healthcare settings (Singhal et al., 2023).

Large language models have recently approached expert-level performance on standardized clinical benchmarks, as illustrated by Med-PaLM 2, which demonstrated substantial improvements in accuracy and clinician-preferred responses on multiple medical QA datasets (Singhal et al., 2025). Such advances highlight LLMs' growing ability to handle complex medical questions, but they also underscore the need for grounded, reliable outputs that align with clinical evidence.

Retrieval-Augmented Generation (RAG) has emerged as a principled strategy to mitigate these risks by conditioning generation on externally retrieved documents (Lewis et al., 2020). By incorporating relevant evidence at inference time, RAG systems aim to improve factual accuracy, traceability, and transparency. In biomedical and clinical domains, retrieval-based augmentation has been applied to clinical guidelines, scientific literature, and electronic health records, often yielding im-

Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)
Vol. 2, pages 112–120

provements in evidence attribution and factual consistency (Xiong et al., 2024). Nevertheless, the prevailing narrative that retrieval universally enhances performance has not been rigorously examined in settings that require higher-order reasoning or contextual abstraction.

This limitation becomes particularly salient in medication-related question answering. Drug leaflets are authoritative regulatory documents that provide structured and legally validated information. However, they are not designed to support complex therapeutic reasoning, differential diagnosis, or integrated clinical decision-making. When retrieval mechanisms introduce partial, excessive, or poorly aligned context, the model may anchor its reasoning to suboptimal evidence. We refer to this phenomenon as *retrieval-induced anchoring bias*. In such cases, responses may remain formally grounded in retrieved passages while still being clinically inappropriate or incomplete. Despite the importance of this issue for patient safety, systematic empirical analyses of retrieval-induced bias in medical QA remain scarce.

The challenge is amplified in Portuguese, which is among the most widely spoken languages globally and the primary language of clinical practice in Brazil. Most studies on medical question answering and retrieval-augmented systems focus predominantly on English-language resources (Singhal et al., 2023). Despite the growing adoption of NLP in healthcare, Portuguese clinical resources remain comparatively scarce. Prior work has focused primarily on foundational components such as domain-specific embeddings (e Oliveira et al., 2019), clinical information extraction from medical records (Da Rocha et al., 2022), and semantically annotated corpora such as SemClinBr (Oliveira et al., 2022).

Specialized biomedical language models have also been developed for Portuguese. For Brazilian Portuguese, BioBERTpt was introduced to support clinical named entity recognition tasks (Schneider et al., 2020), while MediAlbertina is a recent large-scale medical language model for European Portuguese (Nunes et al., 2024). More recent efforts have begun adapting large language models to the medical domain in Portuguese, including studies on fine-tuning strategies and domain transfer (Paiola et al., 2024). In parallel, dedicated benchmarks for Brazilian Portuguese medical NLP have recently emerged, such as the *DrBodeBench* benchmark proposed by (Garcia et al., 2025), highlighting the lack of standardized evaluation resources for clinical reasoning tasks.

However, these resources largely target entity-level understanding or document processing tasks rather than complex clinical reasoning or evidence-grounded question answering. Consequently, publicly available Portuguese benchmarks for clinical QA remain limited, and evaluation frameworks rarely incorporate clinically meaningful dimensions beyond lexical overlap metrics. Standard NLP measures such as Exact Match and F1 do not fully capture groundedness, hallucination risk, or potential clinical harm, dimensions that are increasingly emphasized in trustworthy artificial intelligence research (Rudin, 2019; Ji et al., 2023).

In this work, we conduct a systematic, controlled evaluation of Retrieval-Augmented Generation for medication-related clinical question answering in Portuguese, using Brazilian regulatory drug leaflets as the primary source of authoritative knowledge. Our study is designed to disentangle the role of retrieval across different levels of clinical abstraction. To this end, we evaluate RAG across two complementary, carefully constructed settings. The first setting comprises a controlled, medication-specific QA dataset derived directly from curated regulatory leaflets, with questions tightly aligned with the explicit document content. The second setting consists of a broader clinical reasoning benchmark built from a subset of the Portuguese medical benchmark DrBodeBench (Garcia et al., 2025), which includes questions from Brazilian medical examinations such as Revalida and the FUVEST direct-access residency exam, where medication knowledge must be integrated into complex diagnostic, therapeutic, and decision-making scenarios.

The main contributions of this paper are summarized as follows:

- We introduce a curated Portuguese question answering dataset derived from Brazilian drug leaflets. The dataset includes extensive preprocessing to address crawler-induced extraction noise, structural inconsistencies, and metadata imprecision.

- We construct a complementary clinical QA benchmark by systematically identifying and extracting medication-related questions from the Brazilian medical licensing examination, thereby enabling evaluation in realistic and high-stakes clinical reasoning settings.

- We provide a comprehensive empirical comparison of six modeling strategies, ranging from a base language model without retrieval to multiple retrieval-augmented variants, including fusion-based, hypothetical-document, critique-based, and adaptive approaches.
- We develop a clinically oriented evaluation framework that integrates traditional NLP metrics with groundedness verification, hallucination detection, and explicit clinical risk assessment through an LLM-as-a-judge protocol.
- We offer empirical evidence that Retrieval-Augmented Generation is not universally beneficial in medical question answering and demonstrate that adaptive retrieval control is critical for safe and reliable deployment in complex clinical scenarios.

Overall, our findings contribute to a more nuanced and safety-aware understanding of Retrieval-Augmented Generation in healthcare. We characterize the conditions under which retrieval enhances factual reliability, identify scenarios in which it may introduce reasoning bias, and provide actionable guidance for the responsible development of grounded clinical language models in Portuguese and other underrepresented linguistic contexts.

The remainder of this paper is organized as follows. Section 2 describes the construction and characteristics of the curated drug leaflet dataset and the DrBodeBench-derived medication-focused clinical benchmark. Section 3 details the modeling strategies, including the base language model and its retrieval-augmented variants. Section 4 presents the experimental setup, evaluation protocol, quantitative results, and a comprehensive discussion of clinical safety implications, observed failure modes, and methodological limitations. Finally, Section 5 concludes the paper and outlines directions for future research.

2 Datasets and Benchmark Design

To systematically analyze the effectiveness of Retrieval-Augmented Generation under varying levels of clinical abstraction, we constructed two complementary evaluation benchmarks. The first benchmark isolates document-grounded factual retrieval in a controlled regulatory setting. The second benchmark evaluates retrieval behavior in realistic, high-level clinical reasoning scenarios derived from national medical licensing examinations. This dual design enables a structured investigation of how question specificity and abstraction level modulate retrieval effectiveness.

2.1 Bulário Regulatory Corpus

Our primary knowledge source is the publicly available *Bulário* corpus (Cunha et al., 2018), which contains over 7,000 Brazilian regulatory drug leaflets. These documents constitute the official legal source of information regarding indications, contraindications, dosage, pharmacodynamics, adverse reactions, and drug interactions.

The raw corpus contained significant structural noise introduced by automated crawling during its original compilation. We therefore implemented a multi-stage preprocessing pipeline:

- Removal of duplicated, truncated, or incomplete leaflets;
- Correction of HTML parsing artifacts and formatting inconsistencies;
- Standardization of section headers to enable consistent document segmentation;
- Normalization of metadata fields, including medication names, active substances, and therapeutic classes.

After cleaning, the corpus was indexed using BM25 for lexical retrieval. Preliminary experiments indicated that purely embedding-based semantic retrieval was unstable due to domain-specific terminology and heterogeneous document structure. BM25 provided more reliable alignment with regulatory language and section-level content.

2.2 Medication-Specific QA Benchmark

To construct a controlled evaluation benchmark, we selected 25 widely prescribed medications in Brazil across four therapeutic categories: antibiotics, analgesics and anti-inflammatory agents, antihypertensives, and antidiabetics. These categories were chosen to ensure clinical diversity across infectious, inflammatory, cardiovascular, and metabolic conditions.

For each selected medication, we verified the presence of its corresponding leaflet in the cleaned corpus. We then identified 10 standardized sections consistently present across documents, including indications, dosage, contraindications, warnings, adverse reactions, and drug interactions.

Question generation was performed using a large language model conditioned on the medication name and structured leaflet content. To ensure balanced coverage and section-level control, we generated:

- 4 questions per medication per section;
- 40 questions per section across medications;
- 400 total open-ended questions.

This benchmark, referred to as the **Medication-Specific QA Benchmark**, is explicitly designed to measure factual recall, section-level grounding, and evidence attribution when answers are directly localized within regulatory documents[1].

2.3 Medication-Focused Clinical Benchmark from DrBodeBench

To evaluate retrieval capabilities in higher-level reasoning scenarios, we created a second benchmark derived from the Portuguese medical benchmark DrBodeBench (Garcia et al., 2025). This benchmark aggregates questions from Brazilian medical examinations, including the Revalida and the FUVEST direct-access residency exam. From DrBodeBench, we curated a specific subset of questions that exclusively pertains to medication-related topics.

A large language model was employed to identify examination items in which medication knowledge plays a central role in diagnostic or therapeutic decision-making. Only questions requiring explicit pharmacological integration were retained. The resulting dataset consists of clinically contextualized multiple-choice scenarios that require integration of medication knowledge with patient history, laboratory findings, and clinical reasoning.

In contrast to the controlled leaflet-based benchmark, answers in this dataset are not necessarily localized within a single document section. Instead, they frequently require synthesis across distributed knowledge and contextual interpretation. This property makes the benchmark suitable for analyzing potential retrieval-induced bias in complex reasoning tasks[2].

[1] https://huggingface.co/datasets/recogna-nlp/bulas_qa

[2] https://huggingface.co/datasets/recogna-nlp/drbodebench_medicamentos

3 Methodology

This section describes the modeling framework used to evaluate RAG for medication-related clinical question answering in Portuguese. The study is designed around a single core hypothesis: the effect of retrieval depends on question specificity and on the level of clinical abstraction required. To test this hypothesis, we compare a parametric-only language model against multiple retrieval-augmented variants on two complementary benchmarks, one dominated by localized, document-explicit answers and another requiring higher-level clinical reasoning.

To enable controlled comparisons, we keep the underlying generator fixed across all methods and vary only the retrieval and evidence-integration procedures. Such a controlled setup isolates performance differences attributable to retrieval behavior rather than to model capacity.

3.1 Task Formulation

Let q denote a clinical question and let $\mathcal{D}$ denote the cleaned Bulário collection indexed for retrieval. A retriever R maps q to a ranked list of passages

$$P_K = \{p_1, \ldots, p_K\}, \quad P_K \subset \mathcal{D},$$

where each p_i is a chunk extracted from a drug leaflet, a generator G produces an answer a either from parametric knowledge alone or conditioned on retrieved evidence:

$$a = \begin{cases} G(q) & \text{(parametric-only)}, \\ G(q, P_K) & \text{(retrieval-augmented)}. \end{cases}$$

This formulation supports direct comparisons between parametric-only reasoning and evidence-grounded generation under identical decoding settings. It also allows evaluation across two regimes: medication-specific questions, for which the answer is usually explicitly stated in the leaflet, and complex clinical reasoning questions from the DrBodeBench subset, for which the answer often requires integrating pharmacological knowledge with broader clinical context.

3.2 Evidence Preparation and Retrieval Backbone

Drug leaflets were segmented into section-aware chunks based on standardized headers. Each chunk retained metadata that preserves its provenance, including medication name and section identifier.

This design supports section-level retrieval analysis and enables evaluation of retrieval coverage by leaflet section.

We employ BM25 as the primary retrieval method. In preliminary experiments, purely embedding-based semantic retrieval produced unstable rankings in Portuguese regulatory text, likely due to specialized terminology and heterogeneous formatting. BM25 yielded more robust lexical alignment for this domain and served as the retrieval backbone for all RAG variants.

Unless stated otherwise, all retrieval-augmented methods use a fixed retrieval depth K to ensure comparability across systems.

3.3 Compared Systems

We evaluate six modeling strategies: a parametric-only baseline and five retrieval-augmented methods. All methods use the same generator G and differ only in how they retrieve, filter, or integrate evidence.

3.3.1 Parametric-only Baseline

The **Base** system produces an answer without external evidence:

$$a_{\text{base}} = G(q).$$

This baseline quantifies the performance achievable from parametric knowledge alone and provides a reference point for assessing both gains and degradations introduced by retrieval.

3.3.2 Direct Retrieval-Augmented Methods

RAG-Simple augments generation with the top-K retrieved chunks:

$$P_K = R_{\text{BM25}}(q), \quad a_{\text{simple}} = G(q, P_K).$$

This system tests whether direct evidence injection improves factual accuracy and grounding in medication-specific queries.

RAG-Fusion expands retrieval coverage by issuing multiple query reformulations and aggregating retrieved candidates before selecting the final evidence set P_K. This approach aims to reduce lexical mismatch and improve recall when relevant information is distributed across sections.

RAG-HyDE retrieves evidence using a hypothetical document produced by the generator. The system first generates a synthetic passage $\tilde{d}$ from the question, then retrieves evidence conditioned on $\tilde{d}$, and finally answers using the retrieved context:

$$\tilde{d} = G(q), \;\; P_K = R_{\text{BM25}}(\tilde{d}), \;\; a_{\text{hyde}} = G(q, P_K).$$

HyDE is intended to bridge the gap between the question phrasing and regulatory language, potentially improving retrieval precision.

These direct RAG variants are expected to be most effective when answers are explicitly stated in the documents. In higher-abstraction questions, however, they may introduce partial or poorly prioritized evidence that biases the selection of answers.

3.3.3 Critique-based and Adaptive Retrieval Methods

CRAG introduces an additional validation stage that assesses whether the retrieved context is adequate for the question. When evidence is judged insufficient or misaligned, CRAG triggers a refinement step that updates retrieval and re-generates the answer using improved context. This design targets failure modes in which naive retrieval anchors the model to incomplete evidence in clinically complex scenarios.

Adaptive RAG dynamically decides whether retrieval should be used and controls how much evidence is injected. Formally, a policy $\pi(q)$ selects whether to retrieve and the retrieval depth:

$$(\text{use\_retrieval}, K(q)) = \pi(q).$$

If retrieval is enabled, the system generates $a = G(q, P_{K(q)})$. Otherwise, it returns $a = G(q)$. This adaptive control is designed to mitigate retrieval-induced anchoring bias by reducing reliance on retrieved evidence when the question requires abstraction beyond what leaflets can directly support.

3.4 Design Rationale

The evaluated systems form a progression from parametric-only generation to increasingly controlled retrieval mechanisms. This design supports three empirical tests aligned with our claims: retrieval should improve grounding and factual recall for localized medication questions, naive retrieval may underperform in broader clinical reasoning tasks when evidence is partial or excessive, and adaptive or critique-based strategies should mitigate these failures by controlling when and how retrieval is used.

3.5 Evaluation Overview

All systems are evaluated under identical decoding settings across both benchmarks. We report standard lexical metrics such as Exact Match and F1, retrieval-sensitive metrics such as section-level Recall@K and grounding rate, and clinically oriented dimensions, including hallucination detection and clinical risk assessment, evaluated using the G-Eval framework (Liu et al., 2023), an LLM-based evaluation protocol designed to improve alignment with human judgments. We also report latency to quantify computational overhead introduced by multi-query retrieval and iterative evidence construction.

The full evaluation protocol and results are presented in Section 4.

4 Results and Discussion

The experimental results reveal a consistent, theoretically meaningful pattern: the impact of Retrieval-Augmented Generation is strongly conditioned by question specificity and the level of clinical abstraction required. Retrieval substantially improves performance when answers are explicitly localized in regulatory documents. Its effect becomes more complex in higher-level clinical reasoning scenarios, where uncontrolled evidence injection may interfere with decision-making.

4.1 Results on the Medication-Specific QA Benchmark

Table 1 reports performance on the Medication-Specific QA Benchmark.

Across all retrieval-augmented variants, improvements over the parametric-only baseline are substantial and consistent. The Base model achieves F1 = 0.276 and G-Eval = 0.517, reflecting limited factual precision and reduced clinical coherence even in a restricted domain. All RAG-based methods increase F1 by approximately 0.12 to 0.14 and improve G-Eval by more than 0.20.

These gains confirm that when answers are explicitly stated in regulatory leaflets, retrieval augmentation fulfills its primary objective. It enhances evidence grounding, increases section-level recall, and improves justification quality. Automatic grounding rates and Recall@K values approach ceiling levels across RAG variants, indicating that retrieval coverage is not the principal limiting factor. Performance differences among retrieval methods arise primarily from generation quality and contextual synthesis rather than from evidence availability.

Latency analysis reveals a cost–performance trade-off. RAG-Fusion achieves the highest F1 but nearly doubles inference time. RAG-HyDE attains the highest G-Eval score but incurs the largest computational overhead. CRAG and Adaptive RAG maintain competitive quality while preserving moderate latency. In this extractive setting, simple retrieval offers the most favorable balance between efficiency and performance.

Overall, retrieval is highly effective in document-grounded medication QA, where answers are explicitly localized, and the regulatory structure aligns with the reasoning task.

4.2 Results on the DrBodeBench Clinical Benchmark

A markedly different pattern emerges in the DrBodeBench medication subset, summarized in Table 2.

Direct retrieval strategies do not consistently improve performance in this setting. RAG-Simple and RAG-Fusion slightly underperform the Base model, with degradations of 0.5-1.8 percentage points. RAG-HyDE performs equivalently to the parametric baseline.

Questions in this subset require integrating pharmacological knowledge with patient history, laboratory findings, and therapeutic prioritization. Retrieved leaflet passages often contain accurate but partial information. When such fragments are injected without filtering or abstraction control, the generator may anchor on salient yet incomplete evidence and select suboptimal answer choices.

A comprehensive qualitative analysis presented in Table 3 reveals insights that accuracy alone does not capture. In a representative case involving varicella prophylaxis for an immunosuppressed child, the base model selected the correct answer based on parametric knowledge; however, it lacked explicit supporting evidence, leading the automated judge to classify it as ungrounded.

In contrast, RAG-based methods such as RAG-Simple and RAG-HyDE produced correct answers supported by excerpts from package inserts, which resulted in grounded justifications and reduced clinical risk. On the other hand, RAG-Fusion provided a clinically incorrect answer, despite being grounded in the retrieved content, leading to a classification of grounded and non-hallucinatory, but with high clinical risk.

This case underscores that grounding alone does

Table 1: Performance on the Medication-Specific QA Benchmark. Best results per metric are shown in bold.

Model	F1	G-Eval	Latency (s)
Base (no retrieval)	0.276	0.517	4.41
RAG-Simple	0.411	0.723	4.88
RAG-Fusion	**0.412**	0.723	8.86
RAG-HyDE	0.393	**0.735**	11.64
CRAG	0.393	0.728	5.50
Adaptive RAG	0.401	0.733	5.16

Table 2: Accuracy on the DrBodeBench Medication-Focused Clinical Benchmark. Best result is shown in bold.

Method	Accuracy (%)
Base (no retrieval)	51.85
RAG-Simple	51.36
RAG-Fusion	50.08
RAG-HyDE	51.85
CRAG	**54.25**
Adaptive RAG	52.81

not guarantee clinical accuracy; the retrieval of partially relevant or misleading information can lead the model to make erroneous therapeutic decisions. This behavior exemplifies retrieval-induced anchoring bias, in which retrieved evidence constrains reasoning, even when it does not support the correct answer.

These findings suggest that, for complex clinical reasoning tasks like those in the DrBodeBench, more selective or adaptive retrieval strategies are crucial to mitigate risks. Conversely, direct retrieval remains most effective for straightforward question-answering scenarios, such as those involving package inserts.

Overall, the example underscores the importance of evaluation frameworks that distinguish groundedness from clinical validity.

CRAG achieves the highest accuracy at 54.25%, and Adaptive RAG also surpasses the parametric baseline. Both methods regulate evidence exposure through critique or dynamic control, reducing reliance on misaligned context.

Taken together, the results demonstrate that Retrieval-Augmented Generation is highly effective in document-explicit medication QA but requires adaptive control in higher-abstraction clinical reasoning scenarios to ensure robustness and patient safety.

5 Conclusions and Future Directions

This work presents a systematic evaluation of Retrieval-Augmented Generation for medication-related clinical question answering in Portuguese, grounded in the cleaned Bulário regulatory corpus and validated on a complementary medical licensing benchmark derived from Revalida examinations. By structuring evaluation across two benchmarks that differ in question specificity and abstraction level, the study demonstrates that retrieval effectiveness is not universal. Instead, it depends critically on the specificity of the question and the level of clinical abstraction required.

In document-grounded medication-specific queries, retrieval substantially improves factual recall, evidence grounding, and clinical coherence. When answers are localized within regulatory leaflets, retrieval augmentation enhances reliability and traceability without meaningful performance trade-offs. These findings reinforce the value of RAG architectures in regulatory and extractive clinical information settings, particularly in Portuguese, where high-quality structured benchmarks remain limited.

In contrast, derived from Brazilian medical examination questions included in DrBodeBench, a more nuanced picture emerges. Direct retrieval strategies can slightly degrade performance by introducing partial or misaligned evidence that influences answer selection. This phenomenon, characterized in this study as retrieval-induced anchoring bias, highlights a critical limitation of naive context injection. Grounded responses are not necessarily clinically correct, and the presence of evidence alone does not guarantee safe reasoning.

More advanced retrieval strategies, including critique-based and adaptive approaches, consistently mitigate these limitations. By regulating when and how external evidence is incorporated, these methods improve robustness in clinically complex scenarios and reduce high-risk errors. The

Table 3: Illustrative qualitative example from the DrBodeBench subset.

Method	Grounded	Correct	Clinical Risk
Base	No	Yes	Low
RAG-Simple	Yes	Yes	None
RAG-HyDE	Yes	Yes	None
RAG-Fusion	Yes	No	High
CRAG	Yes	Yes	None

findings suggest that adaptive retrieval control is essential for safe deployment of language models in high-stakes medical environments.

Beyond performance metrics, the study underscores the importance of clinically oriented evaluation frameworks. Standard NLP metrics such as Exact Match and F1 fail to capture safety-relevant differences between systems. Dimensions such as groundedness, hallucination rate, and clinical risk provide complementary insights indispensable to healthcare applications.

Several avenues for future research emerge from these findings. Formal modeling of retrieval-induced anchoring bias could deepen understanding of how exposure to evidence shifts decision boundaries in generative models. Hybrid retrieval mechanisms combining lexical and semantic strategies may improve robustness in morphologically rich languages such as Portuguese. Policy-learning approaches could dynamically calibrate retrieval depth based on clinical risk. Finally, prospective validation in clinical workflows would strengthen the external validity of retrieval-augmented systems.

In summary, Retrieval-Augmented Generation is a powerful but context-dependent approach for clinical question answering. Its benefits are pronounced in document-explicit regulatory domains, yet careful control is required when reasoning extends beyond explicit evidence. For medical AI systems in Portuguese and other underrepresented languages, adaptive retrieval policies are not merely enhancements but foundational requirements for trustworthy and safe deployment.

References

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Alexandre Cunha, Gabriel dos Santos, and Gustavo Paiva Guedes. 2018. Uma análise sobre as bulas de medicamentos no brasil. In *CSBC 2018 - 12º BreSci ()*.

Naila Camila Da Rocha, Abner Macola Pacheco Barbosa, Yaron Oliveira Schnr, Juliana Machado-Rugolo, Luis Gustavo Modelli de Andrade, José Eduardo Corrente, and Liciana Vaz de Arruda Silveira. 2022. Natural language processing to extract information from portuguese-language medical records. *Data*, 8(1):11.

Lucas Emanuel Silva e Oliveira, Yohan Bonescki Gumiel, Arnon Bruno Ventrilho Dos Santos, Lilian Mie Mukai Cintho, Deborah Ribeiro Carvalho, Sadid A Hasan, and Claudia Maria Cabral Moro. 2019. Learning portuguese clinical word embeddings: A multi-specialty and multi-institutional corpus of clinical narratives supporting a downstream biomedical task. In *MedInfo*, pages 123–127.

Gabriel Lino Garcia, João Renato Ribeiro Manesco, Pedro Henrique Paiola, Pedro Henrique Crespan Ribeiro, Ana Lara Alves Garcia, and João Paulo Papa. 2025. A step forward for medical llms in brazilian portuguese: Establishing a benchmark and a strong baseline. In *Proceedings of the 38th IEEE International Symposium on Computer-Based Medical Systems (CBMS 2025)*, Madrid, Spain.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12).

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval:

NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Miguel Nunes, João Boné, João C Ferreira, Pedro Chaves, and Luis B Elvas. 2024. Medialbertina: An european portuguese medical language model. *Computers in Biology and Medicine*, 182:109233.

Lucas Emanuel Silva e Oliveira, Ana Carolina Peters, Adalniza Moura Pucca Da Silva, Caroline Pilatti Gebeluca, Yohan Bonescki Gumiel, Lilian Mie Mukai Cintho, Deborah Ribeiro Carvalho, Sadid Al Hasan, and Claudia Maria Cabral Moro. 2022. Semclinbr-a multi-institutional and multi-specialty semantically annotated corpus for portuguese clinical nlp tasks. *Journal of Biomedical Semantics*, 13(1):13.

Pedro Henrique Paiola, Gabriel Lino Garcia, João Renato Ribeiro Manesco, Mateus Roder, Douglas Rodrigues, and João Paulo Papa. 2024. Adapting llms for the medical domain in portuguese: A study on fine-tuning and model evaluation. *Preprint*, arXiv:2410.00163.

Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215.

Elisa Terumi Rubel Schneider, João Vitor Andrioli de Souza, Julien Knafou, Lucas Emanuel Silva e Oliveira, Jenny Copara, Yohan Bonescki Gumiel, Lucas Ferro Antunes de Oliveira, Emerson Cabrera Paraiso, Douglas Teodoro, and Cláudia Maria Cabral Moro Barra. 2020. Biobertpt-a portuguese neural language model for clinical named entity recognition. In *Proceedings of the 3rd clinical natural language processing workshop*, pages 65–72.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Abubakr Babiker, Nathanael Schärli, Aakanksha Chowdhery, Philip Mansfield, Dina Demner-Fushman, and 13 others. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R. Pfohl, Heather Cole-Lewis, Darlene Neal, Qazi Mamunur Rashid, Mike Schaekermann, Amy Wang, Dev Dash, Jonathan H. Chen, Nigam H. Shah, Sami Lachgar, Philip Andrew Mansfield, and 16 others. 2025. Toward expert-level medical question answering with large language models. *Nature Medicine*, 31(3):943–950.

Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. 2024. Benchmarking retrieval-augmented generation for medicine. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6233–6251, Bangkok, Thailand. Association for Computational Linguistics.

Annotation Guidelines and Challenges for Automatic Simplification of Portuguese Drug Leaflets

Arthur Scalercio[1], **Eduarda Bertotto**[2], **Silvana Jesus**[2], **Maria José Finatto**[2], **Aline Paes**[1]
[1]Institute of Computing, Universidade Federal Fluminense, Niterói, RJ, Brazil,
[2]Institute of Linguistics, Universidade Federal do Rio Grande do Sul, Porto Alegre, RS, Brazil
arthurscalercio@id.uff.br, bertotto.ufrgs@gmail.com, silanjos@gmail.com, mariafinatto@gmail.com, alinepaes@ic.uff.br

Abstract

While most essential medicines have become widely accessible across all social strata in Brazil due to government initiatives and market shifts, a significant barrier remains: the technical complexity of medication leaflets. This pragmatic and linguistic gap hinders patient comprehension of critical risks and benefits. Thus, adapting these texts into plain language patterns is crucial for patient safety and treatment adherence. Large language models have been increasingly effective as practical solutions for text simplification, an important Natural Language Processing (NLP) task that serves as a basis for several other linguistic and computational tasks. However, the scarcity of annotated datasets remains a bottleneck for rigorous evaluation. To bridge this gap, we propose a streamlined pipeline for generating simplified medical leaflets and introduce an initial benchmark dataset of 30 expertly annotated samples. Our results, supported by semantic and morphosyntactic evaluations, demonstrate that the proposed method produces high-quality, simplified content suitable for health applications.

1 Introduction

Brazilian government initiatives, most notably the **Farmácia Popular do Brasil**[1] program, significantly expanded access to essential medicines for treating chronic conditions such as hypertension, diabetes, and asthma (Emmerick et al., 2015). However, widespread access must be matched by health literacy; understanding the risks and correct administration of these drugs is vital for therapeutic success. In this regard, the medication leaflet serves as a critical document, particularly for over-the-counter (OTC) drugs that can be purchased without direct medical supervision. Furthermore, each pharmaceutical company can produce a different package insert with different texts, as long as it respects the minimum mandatory content guidelines of **Anvisa** [Agência Nacional de Vigilância Sanitária], the Brazilian federal agency for drug regulations. Therefore, it is possible to have many different leaflets from different companies for the same medicine and/or substance.

The leaflet includes a section titled "patient information" that provides guidance on how the medication works, its dosage, and its risks. While such a section is theoretically intended for the general public, it is often characterized by dense technical terminology, convoluted sentence structures, and complex formatting. This linguistic barrier creates a gap between access to medicine and its safe use (Roseno, 2024). Given the recent breakthroughs in Large Language Models (Sajjadi Mohammadabadi et al., 2025), Automatic Text Simplification (ATS) (Shardlow, 2014) emerges as a viable solution to bridge this gap, democratizing access to clinical information and ensuring that vital health guidance is intelligible to all social strata.

ATS aims to make written information more accessible by reducing linguistic complexity while preserving meaning (Al-Thanyyan and Azmi, 2021). This task is crucial for a broad audience, including language learners, individuals with cognitive or literacy impairments, and citizens navigating complex institutional or legal documentation (Martínez et al., 2024). Traditionally, ATS research has focused on sentence-level simplification (Al-Thanyyan and Azmi, 2021), where complex sentences are rewritten in isolation. Although valuable, it fails to capture broader *discourse-level phenomena*, such as coherence, conciseness, and referential consistency, which are indispensable for comprehending lengthy or technical documents (Vásquez-Rodríguez et al., 2024). Consequently, *document-level simplification* (DS) has emerged as a more practically relevant yet methodologically challenging task. DS requires the seamless integration of summarization, compression,

[1] https://www.gov.br/saude/pt-br/composicao/sectics/farmacia-popular

Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)
Vol. 2, pages 121–127

and reorganization techniques to sustain global cohesion while ensuring local fluency (Cripwell et al., 2023).

Despite the growing attention toward document-level simplification, most existing research remains restricted to English and relies on a limited set of corpora primarily designed for sentence-level tasks (Ryan et al., 2023). In other languages, particularly Portuguese, the scarcity of structured datasets is a major bottleneck to progress. Pioneering resources such as PorSimples (Aluísio et al., 2008) established a foundation for sentence-level simplification in Brazilian Portuguese; however, there is currently a total absence of *human-annotated* datasets addressing document-level simplification within specialized domains, such as the medical field.

In this work, we present a formal annotation protocol for leaflet simplification and introduce the first human-verified document-level simplification dataset in Portuguese for the medical domain. The dataset was developed using an iterative refinement process: starting from a candidate version initially generated by an LLM, the texts underwent two rigorous rounds of revision. Notably, one of these rounds was conducted by linguists specialized in plain language, ensuring that the simplified versions meet both clinical and accessibility standards. The resulting corpus was then subjected to a two-fold analysis: (i) surface statistics and a readability metric were compared between the original and simple versions and also with an existing simplification dataset, (ii) semantic and morphosyntactic analyses were conducted to automatically assess the quality of the dataset.

This paper makes four main contributions:

1. We propose a simple protocol for annotating simplified medication leaflets by humans.

2. We introduce 30 simplified leaflets, constituting the first human-annotated Portuguese dataset dedicated to document-level simplification in the pharmacological domain. The resource is publicly available for research and educational purposes, adhering to the principles of openness and reproducibility [2].

3. We conduct quantitative analysis of the dataset, comparing its characteristics with those of existing corpora.

4. We assess the quality of the dataset through automatic evaluation using linguistic features.

[2] https://github.com/scalercio/med-simple-docs

2 Annotation Guidelines

This section describes the proposed annotation protocol.

2.1 Starting Point and First Revision

The rules governing the preparation of medication leaflets in Brazil are established in ANVISA Resolution No. 47 of 2009. According to this regulation, patient leaflets must contain three sections: Medication Identification, Patient Information, and Legal Statements. The simplification process was limited to the Patient Information section, which contains nearly all the leaflet's informational content. Under this regulation, the Patient Information section must be organized in a question-and-answer format, consisting of nine questions described in Section 2.2.

Recent studies have increasingly leveraged synthetic data generation to mitigate the scarcity of manually aligned simplification corpora, particularly for low-resource languages and specialized domains (Ghosh et al., 2023; Kaddour and Liu, 2023; Ankinina et al., 2025). Given that Brazilian medication leaflets are typically long documents (Roseno, 2024), we adopted ChatGPT (version 5.2) (OpenAI, 2026) as a starting point for data annotation. This model was selected because of its ability to generate fluent and semantically faithful paraphrases of long documents. The web application was used to generate simplified versions of the texts, and the prompt design was informed by prior work on generating annotated data in the biomedical domain (Attal et al., 2023). The prompt included the following guidelines: (1) Replace technical terms with simpler synonyms; if unavoidable, provide an explanation in parentheses; (2) prefer active voice over passive voice constructions; (3) split long sentences into shorter, simpler ones whenever possible. (4) omit irrelevant sentences; (5) resolve anaphora and ambiguous pronouns; (6) preserve all numerical information, including dosage amounts, frequency, and routes of administration. The automatically generated outputs were subsequently reviewed by one of the non-linguist authors to correct major errors that altered or omitted essential information from the original document. Minor presentation-related adjustments were also performed to ensure clarity and consistency.

2.2 Linguistic Revision

The ATS step applies structural constraints that prioritize shorter sentences and the use of itemized or segmented information to enhance readability. To ensure the leaflets are as clear and accessible as possible, a second review was performed by linguists specialized in plain-language techniques, following experiences of Wives and Finatto (2025) with simplifying texts on cancer for low-educated lay people supported by different LLMs.

To perform the revision, the *primary principle* is to preserve the patient information leaflet in its entirety, in terms of content. Specifically, this involves retaining the nine guiding questions, established by **Anvisa**, about the drug or substance: "1) Para o que este medicamento é indicado?" [What is this medication indicated for?]; "2) Como este medicamento funciona?" [How does this medication work?]; "3) Quando não devo usar este medicamento?" [When should I not use this medication?]; "4) O que devo saber antes de usar este medicamento?" [What should I know before using this medication?] ; "5) Onde, como e por quanto tempo posso guardar este medicamento?" [Where, how, and for how long can I store this medication?]; "6) Como devo usar este medicamento?" [How should I use this medication?]; "7) O que devo fazer quando eu me esquecer de usar este medicamento?" [What should I do if I forget to use this medication?]; "8) Quais os males que este medicamento pode me causar?" [What harm can this medication cause me?]; "9) O que fazer se alguém usar uma quantidade maior do que a indicada deste medicamento?" [What to do if someone takes more than the recommended amount of this medication?].

The revision also aimed to reduce question complexity. For instance, question 8 was reformulated as "Quais problemas este medicamento pode causar?" because *os males* is more complex and the pronoun *me* was considered redundant. Similarly, question 5 was simplified to "Como guardar este medicamento?" [How should this medication be stored?].

The *second principle* was to eliminate excessive information in the simplified version. Such redundancies arise because patient information leaflets in Brazil often contain a lot of topics, as a single leaflet for a given medication or substance may include instructions for treating multiple conditions. For instance, the medication atenolol is indicated for the control of arterial hypertension, angina pectoris, cardiac arrhythmias, myocardial infarction, and early and late treatment after myocardial infarction. Consequently, different patients are required to identify and extract information relevant to their specific condition. Other principles included introducing the lay term before the technical expression, e.g., wheezing (*chiado no peito*), followed by bronchospasm (*broncoespasmo*) in parentheses.

The review further aimed to favor high-frequency lexical choices in Portuguese, such as using "to decrease" (*diminuir*) rather than "to reduce" (*reduzir*), thereby promoting greater clarity and ease of understanding for readers.

3 Drug Leaflets Dataset

This section describes the main characteristics of our initial drug leaflets dataset, which was constructed by simplifying 30 medication leaflets for hypertension drugs included in the Farmácia Popular social program. Ten medications were selected, and for each, three manufacturers were chosen. Appendix A provides the full list of all medications and their respective manufacturers. Appendix B presents an example of the simplification of a drug leaflet. The simplifications were carried out using the protocol described in Section 2.

Surface Statistics Given the absence of other document-level simplification datasets in Portuguese, the dataset was analyzed and compared with the Gov-Lang-BR corpus (Scalercio et al., 2025). Although originally developed for sentence simplification, this dataset also includes multi-sentence examples and texts from the legal domain. The results are shown in Table 1.

		Leaflets	GovLangBR
Original (Source)		30	1,703
Simplified (Target)		30	1,703
Ave. # of sentences	Source	139.0	1.22
	Target	109.7	1.10
Ave. # of words	Source	2,514.5	33.40
	Target	1,571.9	19.35
Ave. # of characters	Source	14,124	181.42
	Target	8,242	104.19

Table 1: Statistics of Leaflets and GovLangBR.

As observed, each instance in our dataset contains substantially more words and sentences than the sentence-level dataset, although both follow the same trend: the simplified version is compressed relative to the original text. While sentence counts decrease modestly (21%), words and characters decrease more aggressively (about 40%).

Document-Level Readability To evaluate the readability of our leaflets dataset, we adopted the Flesch Readability Index (Kincaid et al., 1975), which can be computed without reference data. This metric is grounded in the assumption that shorter words and sentences contribute to easier reading. We employed the version of the formula adapted for Portuguese (Leal et al., 2023). The result is a score, typically between 0 and 100, where higher values indicate easier reading. Table 2 lists the average readability scores of Leaflets and GovLang-BR datasets.

	Leaflets	GovLangBR
Original (Source)	7.3	-12.1
Simplified (Target)	26.5	0.1

Table 2: Comparison of Flesch Readability Index

It can be observed that the gain in average readability for the leaflets dataset is considerably higher than that observed in GovLangBR. Even with a considerable increase in the readability index, the achieved value still indicates that the simplified documents are very difficult to read and are best understood by university graduates. This need for higher literacy levels represents an important challenge. According to the 2024 INAF survey (INAF, 2024), 58% of the Brazilian population has literacy limited to rudimentary and elementary levels, highlighting the strong demand for accessible materials.

The very low readability score of GovLangBR is likely due to the dataset's complexity, characterized by long sentences and technical jargon.

4 Automatic Evaluation

To assess the quality of the dataset, we perform morphosyntactic and semantic analyses.

4.1 Morphosyntactic Evaluation

To evaluate the quality of the simplifications, we employed four reference-free linguistic metrics to assess the examples. They are: (1) Lemma/Token Ratio (LTR) that measures lexical diversity; (2) Ratio of passive to active voice verbs (P/A) to measure more direct constructions; (3) Proportion of adverbial clauses preceding the main clause (AdvL), capturing sentence structure tendencies; and (4) Ratio of fully developed to reduced relative clauses (D/R), reflecting syntactic simplifications. They were developed based on linguistic hypotheses about complexity and are grounded in psycholinguistic research on language processing (Juola, 1998; Gibson, 1998; Charles, 2013; Corrêa et al., 2019). With this approach, we aim to verify that the transformations performed during the simplification process are reflected in the linguistic metrics. Initially, the dataset was annotated morphosyntactically using the UDPipe model, which was trained on a scientific treebank (Straka et al., 2016; de Souza et al., 2021), and, then, we calculated the linguistic metrics.

Dataset	Linguistic Metrics			
	LTR	P/A	AdvL	D/R
Drug Leaflets Set				
Complex	.025	.013	.314	.572
Simple	.032	.003	.482	.647
Simple - reviewed	.029	.003	.427	.697

Table 3: Linguistic Metrics for the Drug Leaflets Dataset

Examining Table 3, with the exception of the LTR metric, all measures follow the expected trend. The slight increase observed in LTR is due to the reduction in the total number of tokens resulting from compression, rather than an increase in the number of lemmas. Regarding the proportion of verbs in the passive and active voices, the first simplification effectivelly eliminates passive constructions, as confirmed by human review. The need to provide explanations and contextualization accounts for the increase in adverbial clauses preceding the main verb in the two simplified versions. Finally, throughout the revision process, the number of reduced relative clauses gradually decreases.

4.2 Semantic Evaluation

To verify semantic consistency between each pair, we employed Sentence-BERT (Reimers and Gurevych, 2019) to compute sentence embeddings. Since the documents consist of multiple sentences, each one was split into smaller chunks, for which individual vector representations were generated. These vectors were then averaged to produce a single embedding per document. We then computed the cosine similarity between the original documents and their linguist-revised simplified versions.

Figure 1 shows the distribution of these scores, along with a kernel density estimation for smoothing the curve. As observed, all the samples in our dataset exhibit a very high semantic similarity to their original documents, which is highly relevant since meaning preservation is a key challenge in document simplification.

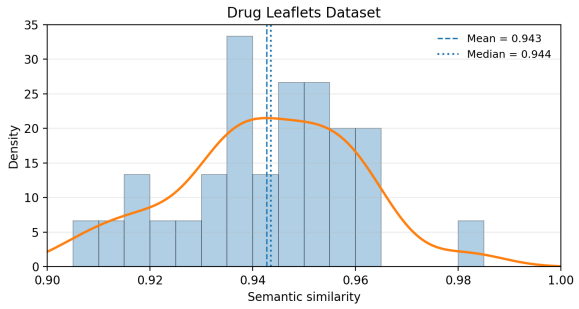


Figure 1: Distribution of Semantic Similarity Scores

5 Conclusions

This work presented a simple annotation protocol for generating simplified medication leaflets, along with an initial dataset of 30 manually simplified leaflets. The main characteristics of the dataset were presented, including an analysis of its readability. Finally, automatic evaluations using linguistic metrics and NLP tools confirmed that the generated simplifications preserve the original content and are simpler than the original leaflets. The main challenge during the annotation process concerns content preservation, given the diversity of existing technical terms and the fact that many spans of text are misspelled and ambiguous. The methodology of human review by specialized linguists ensures that the content is not merely translated by AI, but that it seeks to meet accessibility and clarity standards for the patient. In this regard, we intend to extend the annotation protocol by adding a technical review stage conducted by scholars in the pharmacological field. Other future directions include increasing the number of simplified leaflets, creating an LLM benchmark for document-level simplification, and incorporating different discourse-level metrics into the evaluation framework, such as coherence and cohesion.

Limitations

As mentioned before, the main limitation of our annotation protocol—and consequently of the generated dataset—concerns the lack of a technical review of the simplified leaflets. Without this specialized perspective at all stages, there is a risk that the simplification of the language may simplifying the language may alter or omitt safety. Since preserving the technical and medical content is fundamental for the text simplification task, future work will improve the protocol to include a technical review by pharmaceutical or medical experts, ensuring that, in addition to being easy to read, the text remains clinically accurate and safe. The dataset also presents a limitation in the lack of variety of the selected medications. By selecting only 30 package inserts focused on hypertension medications from the Farmácia Popular program, the study might creates a bias that could compromise the application of the technique in other scenarios. When extending the current dataset, a greater diversity of therapeutic classes will be included to ensure that the simplification protocol works safely and effectively for a wider variety of treatments.

Acknowledgments

This research was supported by CNPq (National Council for Scientific and Technological Development), grant PQ 307088/2023-5, PROBIT and PROBIC FAPERSGS-UFRGS, FAPERJ - *Fundação Carlos Chagas Filho de Amparo à Pesquisa do Estado do Rio de Janeiro*, processes SEI-260003/002930/2024, SEI-260003/000614/2023, and CAPES. We also thank the support of the CNPq National Institutes of Science and Technology, IAIA (grant 406417/2022-9), TILD-IAR (grant 408490/2024-1) and IAPROBEM (grant 408589/2024-8).

References

Suha S Al-Thanyyan and Aqil M Azmi. 2021. Automated text simplification: a survey. *ACM Computing Surveys (CSUR)*, 54(2):1–36.

Sandra M Aluísio, Lucia Specia, Thiago AS Pardo, Erick G Maziero, and Renata PM Fortes. 2008. Towards brazilian portuguese automatic text simplification systems. In *Proc. of the 8th ACM symposium on Document engineering*, pages 240–248.

Tatiana Ankinina, Jan Cegin, Jakub Simko, and Simon Ostermann. 2025. A rigorous evaluation of llm data generation strategies for low-resource languages. *arXiv e-prints*, pages arXiv–2506.

Khalil Attal, Brian Ondov, and Dina Demner-Fushman. 2023. A dataset for plain language adaptation of biomedical abstracts. *Scientific Data*, 10:8.

M Charles. 2013. Active and passive voice in research articles: An interdisciplinary study. *International Journal of Corpus Linguistics*, 18(3):279–318.

Letícia MS Corrêa, Erica dos S Rodrigues, and Renê Forster. 2019. On the processing of object relative clauses. *ExLing 2019*, 25:57.

Liam Cripwell, Joël Legrand, and Claire Gardent. 2023. Document-level planning for text simplification. In

17th Conference of the European Chapter of the Association for Computational Linguistics, pages 993–1006. ACL.

Elvis de Souza, Aline Silveira, Tatiana Cavalcanti, Maria Clara Castro, and Cláudia Freitas. 2021. Petrogold–corpus padrão ouro para o domínio do petróleo. In *Anais do XIII Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 29–38. SBC.

Isabel C. M. Emmerick, Vera L. Luiza, Luiz A. B. Camacho, Christine Vialle-Valentin, and Dennis Ross-Degnan. 2015. Farmácia popular program: Changes in geographic accessibility of medicines during ten years of a medicine subsidy policy in brazil. *Journal of Pharmaceutical Policy and Practice*, 8:10.

Sreyan Ghosh, Chandra Kiran Reddy Evuru, Sonal Kumar, S Ramaneswaran, S Sakshi, Utkarsh Tyagi, and Dinesh Manocha. 2023. DALE: Generative data augmentation for low-resource legal NLP. In *Proc. of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8511–8565. ACL.

Edward Gibson. 1998. Linguistic complexity: Locality of syntactic dependencies. *Cognition*, 68(1):1–76.

INAF. 2024. Indicador de alfabetismo funcional – inaf brasil 2024.

Patrick Juola. 1998. Measuring linguistic complexity: The morphological tier. *Journal of Quantitative Linguistics*, 5(3):206–213.

Jean Kaddour and Qi Liu. 2023. Synthetic data generation in low-resource settings via fine-tuning of large language models. *arXiv preprint arXiv:2310.01119*.

J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical report.

Sidney Evaldo Leal, Magali Sanchez Duran, Carolina Evaristo Scarton, Nathan Siegle Hartmann, and Sandra Maria Aluísio. 2023. Nilc-metrix: assessing the complexity of written and spoken language in brazilian portuguese. *Language Resources & Evaluation*.

Paloma Martínez, Lourdes Moreno, Hiram Ochoa, Alberto Ramos, and Mario Pérez-Enríquez. 2024. A tool suite for cognitive accessibility leveraging easy-to-read resources and simplification strategies. *CEUR-WS.org*.

OpenAI. 2026. Chatgpt-5.2. Accessed: 2026-02-05.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proc. of the 2019 Conference on Empirical Methods in Natural Language Processing*. ACL.

Danilo Alencar Roseno. 2024. *Avaliação da leiturabilidade das bulas dos medicamentos mais comercializados no Brasil e da compatibilidade do uso de medicamentos para dispepsia e constipação durante a amamentação*. Tese (doutorado em assistência farmacêutica), Universidade Federal do Rio Grande do Sul, Porto Alegre, Brazil. Programa de Pós-Graduação em Assistência Farmacêutica. Accessed: 13 Feb. 2026.

Michael J Ryan, Tarek Naous, and Wei Xu. 2023. Revisiting non-english text simplification: A unified multilingual benchmark. In *Proc. of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4898–4927.

Seyed Mahmoud Sajjadi Mohammadabadi, Burak Cem Kara, Can Eyupoglu, Can Uzay, Mehmet Serkan Tosun, and Oktay Karakuş. 2025. A survey of large language models: Evolution, architectures, adaptation, benchmarking, applications, challenges, and societal implications. *Electronics*, 14(18).

Arthur Mariano Rocha De Azevedo Scalercio, Elvis A. De Souza, Maria José Bocorny Finatto, and Aline Paes. 2025. Evaluating LLMs for Portuguese sentence simplification with linguistic insights. In *Proc. of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 24452–24477. ACL.

Matthew Shardlow. 2014. A survey of automated text simplification. *International Journal of Advanced Computer Science and Applications*, 4.

Milan Straka, Jan Hajic, and Jana Straková. 2016. Udpipe: trainable pipeline for processing conll-u files performing tokenization, morphological analysis, pos tagging and parsing. In *Proc. of the 10th International Conference on Language Resources and Evaluation (LREC'16)*, pages 4290–4297.

Laura Vásquez-Rodríguez, Nhung TH Nguyen, Piotr Przybyła, Matthew Shardlow, and Sophia Ananiadou. 2024. Simple is not enough: Document-level text simplification using readability and coherence. *arXiv preprint arXiv:2412.18655*.

Leandro Wives and Maria José Finatto. 2025. Usando llms para simplificar e representar documentos médicos. In *Anais Estendidos do XL Simpósio Brasileiro de Bancos de Dados*, pages 415–425, Porto Alegre, RS, Brasil. SBC.

A Medications and Manufacturers

Table 4 presents the medications and their respective manufacturers for all simplified leaflets.

Medication	Manufacturer
Atenolol 25mg	Germed
Atenolol 25mg	Legrand Pharma
Atenolol 25mg	Sanofi Medley
Besilato de Anlodipino 5mg	Vitamedic
Besilato de Anlodipino 5mg	Novartis
Besilato de Anlodipino 5mg	Achē
Captopril 25mg	Aurobindo
Captopril 25mg	Teuto
Captopril 25mg	Vitamedic
Cloridrato de Propranolol 40mg	1Pharma
Cloridrato de Propranolol 40mg	Globo Pharma
Cloridrato de Propranolol 40mg	Germed
Espironolactona 25mg	EMS
Espironolactona 25mg	Germed
Espironolactona 25mg	Eurofarma
Furosemida 40mg	Neo Química
Furosemida 40mg	Teuto
Furosemida 40mg	Hipolabor
Hidroclorotiazida 25mg	Legrand Pharma
Hidroclorotiazida 25mg	Germed
Hidroclorotiazida 25mg	Teuto
Losartana Potássica 50mg	Cimed
Losartana Potássica 50mg	Legrand Pharma
Losartana Potássica 50mg	Achē
Maleato de Enalapril 10mg	Geolab
Maleato de Enalapril 10mg	Cimed
Maleato de Enalapril 10mg	Biolab
Succinato de Metoprolol 25mg	AstraZeneca
Succinato de Metoprolol 25mg	Neo Química
Succinato de Metoprolol 25mg	Pharlab

Table 4: Medications and Manufacturers of the Simplified Leaflets

B Dataset Example

Below, we present an excerpt from the beginning of the leaflet for enalapril maleate, along with its corresponding simplified version.

Original:

1. PARA QUE ESTE MEDICAMENTO É INDICADO?

Seu médico prescreveu maleato de enalapril para controlar a pressão alta ou melhorar o desempenho do seu coração (tratamento da insuficiência cardíaca). Maleato de enalapril também é usado para a prevenção de insuficiência cardíaca sintomática.

Em muitos pacientes com insuficiência cardíaca que apresentam sintomas, maleato de enalapril retarda a piora da insuficiência cardíaca e reduz a necessidade de internação hospitalar por insuficiência cardíaca. Maleato de enalapril também ajuda alguns desses pacientes a viverem mais. Em muitos pacientes com insuficiência cardíaca em estágio inicial, antes do desenvolvimento dos sintomas, maleato de enalapril ajuda a prevenir o enfraquecimento do desempenho cardíaco e a retardar o aparecimento de sintomas (por exemplo, falta de ar, cansaço após atividades físicas leves, tais como caminhada, ou inchaço dos tornozelos e pés). Esses pacientes poderão precisar de menos hospitalizações por insuficiência cardíaca. Ao tomar maleato de enalapril, alguns pacientes com insuficiência cardíaca podem ter risco mais baixo de sofrer ataque cardíaco (infarto do miocárdio).

Simplified:

1. Para que este medicamento é indicado?

O Maleato de enalapril serve para controlar a pressão alta e para ajudar o coração a funcionar melhor no tratamento da insuficiência cardíaca, quando o coração fica fraco. O Maleato de enalapril também pode ser usado para prevenir os problemas do coração de pessoas que já têm sintomas. Em pessoas que já têm problemas no coração com sintomas, o Maleato de enalapril pode:

- retardar a piora da doença;
- diminuir as internações em hospital;
- aumentar o tempo de vida.

Em pessoas que têm problemas no início e sem sintomas, o Maleato de enalapril pode:

- evitar a perda de força do coração;
- atrasar o aparecimento de sintomas, como falta de ar, cansaço ao caminhar e inchaço nos tornozelos e pés;
- diminuir a necessidade de internações em hospital.

Ao usar Maleato de enalapril, algumas pessoas que já têm insuficiência cardíaca (coração fraco) podem ter menor risco de infarto.

From Annotated Clinical Narratives to Ontology: Structuring Brazilian Portuguese Clinical Data

Fernando Henrique Moura de Oliveira and Cleyton Mário de Oliveira Rodrigues
University of Pernambuco (UPE)
Recife, Pernambuco, Brazil
fhmo@ecomp.poli.br , cleyton.rodrigues@upe.br

Abstract

Clinical NLP for Brazilian Portuguese remains limited by the lack of semantically structured resources that support interoperability and downstream health applications. Although existing corpora provide annotated clinical narratives, their flat annotation schemes restrict semantic expressiveness and alignment with standardized terminologies. In this work, we present a lightweight domain ontology that models clinical entities, contextual qualifiers, and semantic relations in Brazilian Portuguese texts. The ontology is derived from the original corpus annotations and conceptually aligned with standards to enhance interoperability while preserving corpus-specific semantics. This work establishes foundational infrastructure for Portuguese clinical NLP, supporting tasks such as entity normalization, semantic search, and ontology-guided annotation.

1 Introduction

Clinical narratives are a core component of electronic health records (EHRs), encoding diagnoses, symptoms, procedures, and clinical reasoning in free-text form. Converting this unstructured content into structured, machine-interpretable representations remains a central objective of clinical Natural Language Processing (NLP), supporting information extraction, decision support, and interoperability. In Brazilian Portuguese (PT-BR), however, progress is constrained by the scarcity of publicly available annotated EHR datasets, largely due to privacy concerns and the strict requirements of Brazil's data protection law (da Silva and Pazin-Filho, 2025). Although relevant initiatives have emerged, such as the *SemClinBr* corpus with 1,000 annotated notes and over 65,000 entities (Oliveira et al., 2022), a negation-annotated corpus from three hospitals (Dalloux et al., 2021), and task-specific extraction efforts (Kugic et al., 2024), comprehensive resources remain limited, often requiring the development of ad hoc corpora (Sousa et al., 2023). Terminological coverage in PT-BR is also substantially lower than in English in repositories such as UMLS (Névéol et al., 2018), and domain-adapted NLP tools remain underdeveloped, particularly for tasks such as negation scope detection and abbreviation disambiguation (de Souza et al., 2019; da Silva and Pazin-Filho, 2025).

Recent approaches have combined large language models (LLMs) with vector-based retrieval to generate RDF triples aligned with standards such as SNOMED CT[1] , fostering semantic interoperability (Manda, 2025). However, these efforts are typically disease-centered or terminology-driven and rarely focus on ontology construction grounded in annotated clinical corpora. To address this gap, we present ongoing work that transforms the *SemClinBr* corpus into a lightweight domain ontology, adopting a corpus-driven methodology in which classes, relations, and design patterns are derived directly from the annotation schema and empirical linguistic evidence rather than imposed independently of the data.

This paper is organized as follows: Section 2 discusses related ontology work; Section 3 presents the methodology and corpus used; Section 4 reports the results; and Section 5 concludes with final remarks and future work proposals.

2 Related Work

Prior ontology-driven initiatives in Brazilian clinical and biomedical domains, revealing a landscape that is both domain-specific and methodologically heterogeneous. Early efforts such as Santana et al. ((Santana et al., 2011), (Santana et al., 2012)) focused on representing complex biological and epidemiological processes, particularly in neglected tropical diseases and mortality surveillance, emphasizing formal expressiveness through description logics and competency-question validation. Subse-

[1] https://www.snomed.org/

Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)
Vol. 2, pages 128–134

quent works expanded toward applied public health scenarios, notably (Pellison et al., 2020), which leveraged Semantic Web standards to support interoperability across Brazilian tuberculosis information systems.

Other initiatives targeted specialized domains,including mental health decision support (Yamada et al., 2020) and nursing record standardization via ICNP[2] alignment (Yamada et al., 2020)), highlighting ontology alignment and logical consistency as evaluation strategies. More recently, (Bouscarrat et al., 2020)) addressed multilingual disease interoperability using Wikidata[3] and translation-based alignment metrics.

Recent advances in Brazilian Portuguese clinical NLP include (Dutra et al., 2023)'s frame-semantic model for detecting gender-based violence in hospital records and (Da Rocha et al., 2022)'s neural NER system, which extracts clinical entities. Our work complements these approaches by developing an ontology that serves as a bridge and intermediate representation, enabling interoperability with international standards which is a key challenge identified in studies.

3 Material and Methods

3.1 Corpus Ingestion

The initial phase involves acquiring and decompressing the SemClinBr files to access the source XML documents. From each document, four essential data points are extracted: (1) Document ID ("doc_id"): Derived by stripping the extension from the filename to create a unique identifier; (2) Main Text ("main_text"): The complete clinical narrative extracted from the <TEXT> tag; (3) Annotations: Clinical entities (including semantic tags inspired by UMLS[4] semantic groups, character spans, and abbreviations) extracted from the <TAGS> block and stored as dictionaries containing their ID, type, and text position; and (4) Relations: Dependencies between annotations (e.g., associated_with, negation_of) parsed from the <RELATIONS> block, linking annotation pairs to their specific relationship type.

3.2 Ontology Construction

The ontological schema defines a foundational framework of 14 high-level abstract classes via the `owl:Class` construct. These classes, such as `ClinicalEpisode`, `BodyStructure`, `Substance`, and `ClinicalFinding`, provide the structural backbone for the corpus data. Relationships between these classes are formalized through `owl:ObjectProperty` declarations, utilizing `rdfs:domain` and `rdfs:range` constraints to ensure semantic integrity. For example, the `hasFinding` property maps a `ClinicalEpisode` to a `Finding`, thereby enforcing domain-specific logic and ensuring that clinical observations are correctly contextualized within specific patient encounters.

A critical feature of this design is the explicit modeling of semantic associations and negations. The `associated_with` property serves as a general-purpose link between related entities when specific sub-properties are not defined. Furthermore, to support accurate clinical reasoning, the ontology implements a negation mechanism via a dedicated `negation_of` property. For instance, the absence of a clinical sign is represented by linking the individual representing that sign to a negation individual (e.g., `Concept_sem`, signifying "without"), ensuring the system distinguishes between the presence and absence of clinical findings.

3.3 Ontology Population

The ontology was automatically populated from the annotated SemClinBr corpus using a structured instantiation pipeline implemented in Python with Owlready2 and executed in Google Colab to ensure reproducibility.

Each unique annotated span was instantiated as an `owl:NamedIndividual` following the naming convention `Concept_[Portuguese_term]`, preserving traceability to the source text. Individuals were explicitly typed using `rdf:type`. The population workflow comprised four stages: (1) ontology loading, (2) URI sanitization, (3) individual instantiation, and (4) semantic relation assertion.

Tags were assigned to the most specific ontology class through a prefix-based mapping derived from SemClinBr tags. When no direct mapping was available, the system defaulted to the superclass *Finding*, ensuring complete coverage without interrupting execution. Each clinical document (`doc_id`) was instantiated as an individual of class *ClinicalEpisode*. Sanitized identifiers were preserved via the `hasID` data property.

Episode individuals were linked to concept instances through semantically appropriate object

[2]https://www.icn.ch/icnp-browser
[3]https://www.wikidata.org/wiki/Wikidata:Main$_{Page}$
[4]https://www.nlm.nih.gov/research/umls/index.html

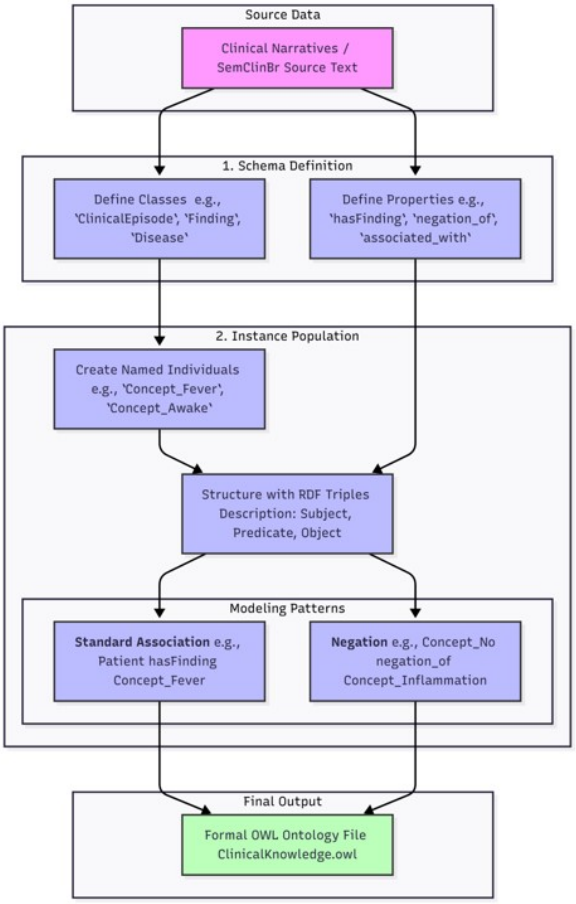


Figure 1: Overview of the corpus-driven workflow for ontology population and semantic relation assertion.

properties , such as `hasFinding`, `hasDisease`, and `hasMedication`. Two categories of relations were asserted. Episode-to-concept relations connect clinical episodes to findings, diseases, procedures, and other entities. Concept-to-concept relations encode semantic interactions such as `associated_with`, and `negation_of`.

Post-population validation included: (i) verification of individual counts, (ii) enumeration of object and data properties, (iii) logical consistency checking using the HermiT reasoner (Horrocks et al., 2012) in Protégé (Musen, 2015), and SPARQL[5] queries.

Figure 1 summarizes the transformation pipeline from annotated narratives to the final OWL[6] ontology (`ClinicalKnowledge.owl`).

[5]https://www.w3.org/TR/sparql11-query/
[6]https://en.wikipedia.org/wiki/Owl

Table 1: Top 10 Annotation Types Frequency

Type	Freq.
Finding	4304
Sign or Symptom	3537
Quantitative Concept	3447
Health Care Activity	2155
Therapeutic or Preventive Procedure	1805
Lab or Test Result	1518
Negation	1514
Disease or Syndrome	1345
Temporal Concept	1230
Abbreviation	1171

Table 2: Frequency of Relation Types

Relation	Freq.
associated_with	9852
negation_of	1606

4 Results and Discussion

4.1 Corpus Analysis

Table1 shows that the corpus is dominated by descriptive clinical entities, particularly *Finding* and *Sign or Symptom*, followed by measurement- and procedure-related categories (e.g., *Quantitative Concept*, *Therapeutic or Preventive Procedure*). Linguistic types such as *Negation* and *Abbreviation* further emphasize contextual and polarity modeling. As shown in Table 2, `associated_with` is the predominant relation, indicating dense semantic connectivity, while `negation_of` encodes clinically relevant polarity.

Although inspired by UMLS semantic groups, the hierarchy lacks explicit description-logic constraints and formal alignment with SNOMED CT. Several classes participate in polyhierarchical structures, and linguistic constructs (e.g., *Negation*, *Abbreviation*, *Temporal Concept*) are modeled alongside biomedical entities. This mixing of abstraction levels introduces semantic overlap and weakens subsumption clarity. Such structural characteristics limit automated reasoning and logical consistency. To support ontology construction and semantic interoperability, SemClinBr requires reorganization toward a higher-level hierarchy more closely aligned with SNOMED CT.

4.2 Ontology Analysis

Table 3 summarizes the global ontology metrics. The ontology was implemented using Protegé (Musen, 2015) and comprises 57 classes, 17,663 individuals, 23 object properties, and 7 data prop-

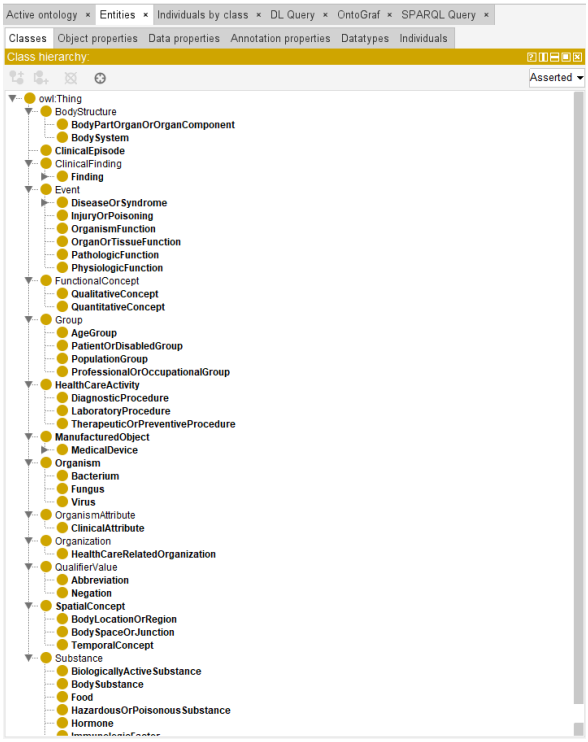


Figure 2: Protegé Ontology Class Hierarchy.

Table 3: Global Ontology Statistics

Metric	Value
Axiom	87,634
Class Count	57
Individual Count	17,663
Object Property Count	23
Data Property Count	7
Hierarchy Maximum Depth	6
Clinical Episodes	1,000

erties. The class hierarchy reaches a maximum depth of 6 levels, indicating moderate taxonomic complexity.

All individuals were typed, following two dominant naming patterns: *Concept_* and *ClinicalEpisode_*. The most populated semantic categories are *Abbreviation*, *Finding*, and *SignOrSymptom*, followed by *LaboratoryOrTestResult* and *QuantitativeConcept*. The ontology encodes 1,000 clinical episodes with an average of 16 connections per episode.

The presence of *ClinicalAttribute* and *Negation* as classes in Figure 2 stems from the SemClinBr annotation schema, which treats these as primary semantic tags. Our current lightweight ontology prioritizes a high-fidelity mapping of the corpus's entity types and relation types. We acknowledge that in formal ontology engineering, these should be refactored into properties.

Duplicate instances were detected due to case variation, accent normalization, token segmentation, and word-order permutation (e.g., "paciente", "PACIENTE", "Paciente"; "hemodinamicamente estável" vs. "estável hemodinamicamente"; "sistema fechado de aspiração" vs. "sistema de aspiração fechado"). These surface variations represent semantically equivalent entities and artificially inflate instance counts suggesting the need for lexical normalization and canonicalization strategies.

4.3 Consistence and Validation

Reasoning over the ontology confirmed logical consistency, with no detected contradictions, and enabled the inference of additional implicit relations. SPARQL queries executed over the inferred ontology revealed a coherent and densely connected structural backbone, consistent with the rich and narrative-driven nature of clinical records.

The most frequent and structurally dominant relation is associated_with, which appears in thousands of triples, reflecting the broad associative nature of clinical documentation where findings, procedures, substances, and other concepts are frequently co-occurring or contextually linked within the same episode. In contrast, semantically more precise and critical relations such as negation_of appear in a smaller but highly meaningful subset (e.g., 27 results in the negation-focused query), highlighting negation patterns that are essential for accurate clinical interpretation and modeling refinement.

Other targeted relations (*hasFinding*, *hasDisease*, etc.) show consistent and expected usage patterns, with results typically ranging from dozens to hundreds of triples depending on the specificity of the query. These patterns confirm that the ontology captures the multifaceted and relational character of real-world clinical knowledge, while also exposing opportunities for further refinement, particularly in strengthening negation modeling and reducing overly generic associated_with links where more precise semantics (e.g., causes, treats, located_in) could be introduced.

Screenshots from Protegé queries (Figures 3 to 5) are provided in the Appendix section for reference.

5 Conclusion and Future Work

In summary, the SemClinBr-derived ontology provides a valuable intermediate representation that bridges clinical NLP annotations and formal ontology engineering. Although lightweight, the ontology is well-structured, as demonstrated by

SPARQL queries that reveal its capacity to model complex clinical relations and support disambiguation. For example, the explicit use of properties such as negation_of allows the system to distinguish between the presence and absence of clinical findings, while the associated_with relation captures contextual links between entities. These features, combined with the ontology's ability to represent and query semantic relations, facilitate more precise clinical interpretation and reduce ambiguity in narrative data.

Although this work establishes a foundational infrastructure for Brazilian Portuguese clinical NLP, we recognize that the research is currently in an initial stage. Our future roadmap includes the implementation of lexical normalization and canonicalization strategies to handle surface variations (e.g., case and accentuation) that currently inflate instance counts. Furthermore, upcoming developments will focus on the explicit modeling of disease–symptom relations and full alignment with global standards, such as ICD[7] and SNOMED CT, to ensure semantic interoperability across diverse healthcare applications.

To transition this resource from a valuable intermediate representation into a formal domain ontology, future work will focus on concept consolidation and lexical normalization to resolve the current overlap between domain and linguistic layers. These efforts will include the implementation of stronger logical axiomatization such as disjointness domain/range constraints, and the explicit modeling of disease–symptom relations. Ultimately, alignment with standards will ensure that the ontology provides a rigorous framework for clinical reasoning and automated inference for Brazilian Portuguese healthcare data.

6 Acknowledgments

We thank the colleagues at the Health Artificial Intelligence Lab (HAILab-PUCPR)[8] for granting access to their Portuguese clinical corpus, which made this study possible. Due to privacy and data protection considerations, the corpus cannot be publicly shared.

References

Léo Bouscarrat, Antoine Bonnefoy, Cécile Capponi, and Carlos Ramisch. 2020. Multilingual enrichment of disease biomedical ontologies. In *Proceedings of the lrec 2020 workshop on multilingual biomedical text processing (multilingualbio 2020)*, pages 21–28.

Naila Camila Da Rocha, Abner Macola Pacheco Barbosa, Yaron Oliveira Schnr, Juliana Machado-Rugolo, Luis Gustavo Modelli de Andrade, José Eduardo Corrente, and Liciana Vaz de Arruda Silveira. 2022. Natural language processing to extract information from portuguese-language medical records. *Data*, 8(1):11.

Rildo Pinto da Silva and Antonio Pazin-Filho. 2025. Dataset of anonymized discharge summaries of sepsis patients from a brazilian tertiary hospital for nlp applications. *Data in Brief*, 61:111804.

Clément Dalloux, Vincent Claveau, Natalia Grabar, Lucas Emanuel Silva Oliveira, Claudia Maria Cabral Moro, Yohan Bonescki Gumiel, and Deborah Ribeiro Carvalho. 2021. Supervised learning for the detection of negation and of its scope in french and brazilian portuguese biomedical corpora. *Natural Language Engineering*, 27(2):181–201.

Joao Vitor Andrioli de Souza, Yohan Bonescki Gumiel, Lucas Emanuel Silva, Claudia Maria Cabral Moro, and 1 others. 2019. Named entity recognition for clinical portuguese corpus with conditional random fields and semantic groups. In *Simpósio Brasileiro de Computação Aplicada à Saúde (SBCAS)*, pages 318–323. SBC.

Lívia Dutra, Arthur Lorenzi, Lorena Larré, Frederico Belcavello, Ely Edison da Silva Matos, Amanda Pestana, Kenneth Brown, Mariana Gonalves, Victor Herbst, Sofia Reinach, and 1 others. 2023. Building a frame-semantic model of the healthcare domain: Towards the identification of gender-based violence in public health data. In *Proceedings of the 14th Brazilian Symposium in Information and Human Language Technology*, pages 347–355.

Ian Horrocks, Boris Motik, and Zhe Wang. 2012. The hermit owl reasoner. *Journal of Automated Reasoning*, 858(3).

Amila Kugic, Stefan Schulz, and Markus Kreuzthaler. 2024. Disambiguation of acronyms in clinical narratives with large language models. *Journal of the American Medical Informatics Association*, 31(9):2040–2046.

Prashanti Manda. 2025. Large language models in bio-ontology research: A review. *Bioengineering*, 12(11):1260.

Mark A Musen. 2015. The protégé project: a look back and a look forward. *AI matters*, 1(4):4–12.

Aurélie Névéol, Hercules Dalianis, Sumithra Velupillai, Guergana Savova, and Pierre Zweigenbaum. 2018. Clinical natural language processing in languages other than english: opportunities and challenges. *Journal of biomedical semantics*, 9(1):12.

[7] https://icd.who.int/en/

[8] https://github.com/HAILab-PUCPR

Lucas Emanuel Silva e Oliveira, Ana Carolina Peters, Adalniza Moura Pucca Da Silva, Caroline Pilatti Gebeluca, Yohan Bonescki Gumiel, Lilian Mie Mukai Cintho, Deborah Ribeiro Carvalho, Sadid Al Hasan, and Claudia Maria Cabral Moro. 2022. Semclinbr-a multi-institutional and multi-specialty semantically annotated corpus for portuguese clinical nlp tasks. *Journal of Biomedical Semantics*, 13(1):13.

Felipe Carvalho Pellison, Rui Pedro Charters Lopes Rijo, Vinicius Costa Lima, Nathalia Yukie Crepaldi, Filipe Andrade Bernardi, Rafael Mello Galliez, Afrânio Kritski, Kumar Abhishek, and Domingos Alves. 2020. Data integration in the brazilian public health system for tuberculosis: use of the semantic web to establish interoperability. *JMIR Medical Informatics*, 8(7):e17176.

Filipe Santana, Fred Freitas, Roberta Fernandes, Zulma Medeiros, and Daniel Schober. 2012. Towards an ontological representation of morbidity and mortality in description logics. *Journal of biomedical semantics*, 3(Suppl 2):S7.

Filipe Santana, Daniel Schober, Zulma Medeiros, Fred Freitas, and Stefan Schulz. 2011. Ontology patterns for tabular representations of biomedical knowledge on neglected tropical diseases. *Bioinformatics*, 27(13):i349–i356.

Hugo Sousa, Alipio Mario Jorge, Arian Pasquali, Catarina Santos, and Mario Lopes. 2023. A biomedical entity extraction pipeline for oncology health records in portuguese. In *Proceedings of the 38th ACM/SIGAPP Symposium on Applied Computing*, pages 950–956.

Diego Bettiol Yamada, Filipe Andrade Bernardi, Newton Shydeo Brandão Miyoshi, Inácia Bezerra de Lima, André Luiz Teixeira Vinci, Vinicius Tohoru Yoshiura, and Domingos Alves. 2020. Ontology-based inference for supporting clinical decisions in mental health. In *International Conference on Computational Science*, pages 363–375. Springer.

A Appendix

Snap SPARQL Query:

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>


PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>

SELECT ?c1 ?c2
WHERE {
  ?c1 <http://www.semantic-clinical-records.org/ontology#associated_with> ?c2 .
}
ORDER BY ?c1
```

Execute

?c1	?c2
<http://www.semantic-clinical-records.org/ontology#Concept_01_cp>	<http://www.semantic-clinical-records.org/ontology#Concept_ap_s_uma_semana>
<http://www.semantic-clinical-records.org/ontology#Concept_01_cp>	<http://www.semantic-clinical-records.org/ontology#Concept_aumenta>
<http://www.semantic-clinical-records.org/ontology#Concept_01cp>	<http://www.semantic-clinical-records.org/ontology#Concept_marevan>
<http://www.semantic-clinical-records.org/ontology#Concept_01cp>	<http://www.semantic-clinical-records.org/ontology#Concept_domingos>
<http://www.semantic-clinical-records.org/ontology#Concept_02_EM_N_VOA_CONT_NUA>	<http://www.semantic-clinical-records.org/ontology#Concept_SPO2_97>
<http://www.semantic-clinical-records.org/ontology#Concept_02_MG>	<http://www.semantic-clinical-records.org/ontology#Concept_RISPERIDONA>

3108 results

Figure 3: SPARQL results: associated_with relations between clinical concepts (3,108 triples).

Snap SPARQL Query:

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>

PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>

SELECT ?neg ?target
WHERE {
  ?neg rdf:type <http://www.semantic-clinical-records.org/ontology#Negation> .
  ?neg <http://www.semantic-clinical-records.org/ontology#negation_of> ?target .
}
ORDER BY ?neg
```

Execute

?neg	?target
<http://www.semantic-clinical-records.org/ontology#Concept_AUSENCIA>	<http://www.semantic-clinical-records.org/ontology#Concept_SINAIS_DE_INFEC_O>
<http://www.semantic-clinical-records.org/ontology#Concept_AUSENTES>	<http://www.semantic-clinical-records.org/ontology#Concept_ELIMINA_ES_INTESTINAIS>
<http://www.semantic-clinical-records.org/ontology#Concept_NEGA>	<http://www.semantic-clinical-records.org/ontology#Concept_DISPNEIA>
<http://www.semantic-clinical-records.org/ontology#Concept_NEfa>	<http://www.semantic-clinical-records.org/ontology#Concept_v_mitos>
<http://www.semantic-clinical-records.org/ontology#Concept_NEfa>	<http://www.semantic-clinical-records.org/ontology#Concept_perda_de_sconsci_ncia>
<http://www.semantic-clinical-records.org/ontology#Concept_NUNCA>	<http://www.semantic-clinical-records.org/ontology#Concept_COLICA_BILIAR>

27 results

Figure 4: SPARQL results: negation_of assertions (27 triples).

Snap SPARQL Query:

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>


SELECT ?episode ?food
WHERE {
  ?episode rdf:type <http://www.semantic-clinical-records.org/ontology#ClinicalEpisode> .
  ?episode <http://www.semantic-clinical-records.org/ontology#hasFinding> ?food .
  ?food rdf:type <http://www.semantic-clinical-records.org/ontology#Food> .
}
```

Execute

?episode	?food
<http://www.semantic-clinical-records.org/ontology#ClinicalEpisode_9441>	<http://www.semantic-clinical-records.org/ontology#Concept_presunto>
<http://www.semantic-clinical-records.org/ontology#ClinicalEpisode_9441>	<http://www.semantic-clinical-records.org/ontology#Concept_salgado>
<http://www.semantic-clinical-records.org/ontology#ClinicalEpisode_9441>	<http://www.semantic-clinical-records.org/ontology#Concept_salada>
<http://www.semantic-clinical-records.org/ontology#ClinicalEpisode_9441>	<http://www.semantic-clinical-records.org/ontology#Concept_queijo>
<http://www.semantic-clinical-records.org/ontology#ClinicalEpisode_9441>	<http://www.semantic-clinical-records.org/ontology#Concept_arroz>
<http://www.semantic-clinical-records.org/ontology#ClinicalEpisode_9441>	<http://www.semantic-clinical-records.org/ontology#Concept_salsicha>

36 results

Figure 5: SPARQL results: 36 Clinical episodes associated with Food findings (hasFinding → Food).

The visible and the latent linguistic clues of mental health in Brazilian Portuguese textual posts

Rodrigo Wilkens[1], Helena Caseli[2], Vânia Neris[2], Aline Villavicencio[1]
[1]University of Exeter,
[2]Universidade Federal de São Carlos (UFSCar),
Correspondence: r.wilkens@exeter.ac.uk

Abstract

Depressive symptomatology may be reflected in the language used by people with possible depressive profiles (PDP). This paper investigates to what extent symptoms of depression are manifested in Brazilian Portuguese narrative texts, and whether these can be used to identify relevant linguistic clues related to PDP. Moreover, the relation between these symptoms and PDP is explored, characterising the lexical, syntactic, and psycholinguistic aspects of texts produced by PDP. We found that texts associated with PDPs differed in some of these characteristics from non-PDP texts. The interactions between symptoms and PDP can also shed light on patterns of communication differentiation and the relationship between them. The results of this paper can help to characterise and understand the linguistic indicators that can be used to train more bespoke and accurate language models.

1 Introduction

Depression is a condition that affects 4% of the global population (Global Burden of Disease Collaborative Network, 2025). In Brazil, the prevalence of depressive disorder is even higher, affecting almost 17.4% of the adult population (Errazuriza et al., 2023). Depression is a clinically significant psychological disorder that causes a range of impairments as insomnia, memory loss, loss of pleasure, and a resulting decline in quality of life. In its most severe forms, depression may result in death by suicide. Despite its widespread prevalence and serious consequences, only about a quarter of individuals with mental disorders receive appropriate care (World Health Organization, 2021).

These challenges motivate the creation of complementary depression detection techniques, with broader reach, to enable the allocation of mental health resources to individuals and populations that would otherwise have no or poor access to them. These techniques may be used to analyse user-generated content published on social media platforms and other text-based reporting tools, such as personal diaries. Written expressions — whether shared on social media or recorded in personal diaries — can serve as valuable early indicators of depression. Recognizing these signs enables for timely prevention and prompt initiation of mental health care. Indeed, Natural Language Processing (NLP) can be used to support depression detection, as texts produced by individuals suffering from depression may contain clear signs of the condition — such as explicit mentions of symptoms (Yazdavar et al., 2017; Yadav et al., 2020; Mendes and Caseli, 2024) and self-declarations of diagnosis/ongoing treatment (Santos et al., 2023) — as well as more subtle clues like language style.

Although prior research demonstrates the utility of NLP for identifying depression (Yazdavar et al., 2017; Yadav et al., 2020), cultural factors can significantly shape how individuals express themselves. Therefore, it is critical to investigate these signs in different languages before starting NLP tools development. Investigating linguistic clues across diverse corpora is essential for advancing accurate computational technologies that can automatically process and understand the language linked to Possible Depressive Profile (PDP) in various cultural contexts. To do so for Brazilian Portuguese texts, we investigated some specific phenomena in this language (e.g. null subjects/person morphology) and also the combination of lexical, syntactic, and psycholinguistic features. This work is carried out in the scope of the AIM-Health[1] research project.

This paper investigates how symptoms of depression are manifested in Brazilian Portuguese narrative texts to identify relevant linguistic markers. The data consists of anonymous Facebook posts from the Amive corpus (Mendes and Caseli,

[1] https://www.aim-health.ufscar.br/

Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)
Vol. 2, pages 135–147

2024) annotated at symptom-level (18 symptoms of depression) and user-level (indicating PDP). We address the following three research questions:

RQ1 Can the overt linguistic clues in narratives be used to identify cases of PDP?

RQ2 What are the linguistic characteristics of the symptoms presented in the Amive corpus?

RQ3 What are the symptoms linked to PDP?

These research questions lay the groundwork for using this corpus in Portuguese-language automatic processing of depression and mental health contexts. They also provide insights that can guide the analysis of other health-related corpora.

Our research approach involved exploring the lexical, syntactic, and psycholinguistic characteristics of the annotated texts. The results indicate clear differences between texts expressing signs of depression and those that do not. Accordingly, we present and discuss these characteristics that may serve as valuable features for training NLP models.

2 Related Work

De Choudhury et al. 2013 were one of the first to investigate the use of social media text for the automatic prediction of depression. They investigated the task of user-level depression classification based on features like emotional content from LIWC (Tausczik and Pennebaker, 2010) categories and activation/arousal from ANEW (Bradley and Lang, 1999), and depression related terms, language style (syntactic features) among others.

Ji et al. 2018 investigated Reddit and Twitter posts focusing on suicidal ideation. They also applied LIWC (Pennebaker et al., 2015) to analyse suicide-related texts and suicide-free posts. They point out differences in language style with a more direct and aggressive language for Twitter users, who tend to make direct mention of their suicidal ideation and somatic symptoms, while Reddit users employ a more indirect speech, contextualizing their feelings and talking about other subjects like finance, social life, and family. The authors also observed similarities between the two datasets, such as the increased frequency of first-person usage in suicidal users. The authors also used LDA (Blei et al., 2003) to topic analysis from posts containing suicidal ideation and summarized the results in three categories: internal factors (words expressing people's feelings, intentions, and desires like "know", "want", "feel" and "like", and "hope"), external social factors (topics containing words such as "money" and "working", "friend", "school", "surgery", "crisis", and "accident"), and the mixed internal/external factors.

Trifu et al. 2024 investigated the linguistic markers for major depressive disorder using a Romanian version of LIWC[2]. They analysed word categories related to: part-of-speech tags, affect words (e.g. positive, negative, anxiety), social words (e.g. family, friends), cognitive process (e.g. insight, causations, certainty), perceptual process (e.g. see, hear, feel), biological process (e.g. body, health, sexual), drives (e.g. affiliation, achievement, power), time-oriented (past, present, future), personal concern (e.g. work, leisure, money), and informal language. They concluded that the language used by depressive patients is significantly different from people without mood disorders in both form – such as the prevalence of short sentences, impersonal pronouns, first-person pronouns in plural form, conjunctions, auxiliary verbs and negations – and content – words indicating negative affects, anxiety in contrast to words indicating positive affects.

For the Portuguese language, a few studies investigate language technologies in mental health domain. Santos et al. 2023 delivered SetembroBR, a large corpus for user-level depression and anxiety classification composed of posts from self-declared depressed Brazilian Twitter users. Mann et al. 2020 developed multimodal classifiers for user-level depression detection based on textual and visual information from Brazilian college students' Instagram posts and their answers to a BDI (Beck Depression Inventory) questionnaire. Text features were TF-IDF, BoW, pretrained FastText and ELMo embeddings, and visual features were ResNet/ResNext embeddings, fused in a fully connected multimodal layer. LIWC categories were also used in traditional machine learning models. Mann et al. 2025 focused on topic-detection in Instagram, Twitter and Reddit datasets in Brazilian Portuguese in a user-level classification approach. They performed experiments with LDA-based models, BERTopic (Grootendorst, 2022) and large language models (LLMs) and found that the LLM-based method yielded a broader and more varied range of topics than conventional techniques. They found an overlap in the topics discussed by users with and

[2]Available at: https://www.liwc.app/.

without depression, which highlights the difficulty of classifying a user based solely on their topics.

Casani et al. 2021 and Mendes and Caseli 2024 investigated multiclass depression symptom classification in Portuguese Twitter and Facebook posts, respectively. In both cases, the corpora were manually labelled by mental health experts. In (Casani et al., 2021), only 3 symptoms categories were considered – psychological, physiological, and behavioural – and a dataset of 2,008 annotated posts was generated. In (Mendes and Caseli, 2024), 21 signals of depression (18 symptoms and 3 other signals) were defined and a dataset of 2,304 annotated symptoms was generated. However, none of these works for Portuguese language brings a linguistic analysis of the built datasets.

In this paper we investigate the linguistic clues and the symptoms relationship in the same corpus of (Mendes and Caseli, 2024): the Amive corpus.

3 Materials and Methods

3.1 The Amive Corpus

A key methodological aspect of this study concerns corpus selection. Identifying naturally occurring expressions of depression in large-scale online data is challenging due to the relatively low prevalence of explicit references to depressive symptoms in general discourse. Random sampling of social media content would therefore yield an extremely low signal-to-noise ratio. To address this issue, we rely on the Amive corpus (Mendes and Caseli, 2024), which was collected from public Facebook pages dedicated to sharing posts written by Brazilian college students. Although this targeted setting introduces potential sampling limitations (e.g., demographic concentration and platform-specific discourse norms), it substantially increases the likelihood of retrieving content relevant to depressive experiences. Given the volume of available material, further filtering was necessary to make human annotation feasible. According to the authors, posts were collected based on keywords related to depression ("suicide", "depression", "cut myself", "will to live", "kill myself" and "want to die") and publication date from January 1st, 2012 to December 31st, 2021.

After corpus construction, it was annotated[3] in two levels: user-level and sentence-level. In the user-level, each post was annotated as a Possible Depressive Profile (PDP) or not (non-PDP). In the sentence-level, texts spans were annotated regarding 18 symptoms of depression: (1) *Agitation/Restlessness*, (2) *Attention/memory deficit*, (3) *Alteration in weight/eating habits*, (4) *Sleep disorder*, (5) *Physical symptom*, (6) *Difficulty in decision-making*, (7) *Feeling of emptiness*, (8) *Loss/Diminishment of pleasure/libido*, (9) *Feeling of guilt*, (10) *Irritation/Aggressiveness*, (11) *Alteration in efficiency/functionality*, (12) *Tiredness/Discouragement/Fatigue*, (13) *Despair*, (14) *Worry/Fear/Anxiety*, (15) *Feeling of worthlessness/Low self-esteem*, (16) *Suicide/Self-extermination*, (17) *Helplessness/Social harm/Loneliness*, and (18) *Sadness/Depressed mood*.

For this study, the original annotation was revised to standardize the boundaries of the text spans and to certify that all annotated segments were present in the final corpus. Thus, the corpus under analysis in this paper has 604 textual posts, 336 of them annotated as PDP and 268 as non-PDP. It is worth mention that it is possible to have annotated symptoms in non-PDP posts. Table 1 shows the amount of instances annotated for each symptom. Although notes may extend beyond the sentence boundary, this rarely occurs.[4] Table 2 shows the total and average number of sentences, tokens, types, symptoms, and unique symptoms per text.

Table 3 brings some instances from the original corpus translated to English and slightly modified to anonymization purpose.

3.2 Linguistic characterization of symptoms and PDP

To address our first two research questions, we examined the lexical, syntactic, and psycholinguistic characteristics of the annotated texts summarized in Table 4. For both symptom-annotated spans and texts labelled as PDP or non-PDP, we computed the number of tokens, types, Type–Token Ratio (TTR), span length, mean syllables per word, and the Flesch readability score adapted for Brazilian Portuguese (Martins et al., 1996). These measures were selected to capture lexical richness, verbosity,

[3]According to Mendes and Caseli 2024, annotation was performed by four students from psychology, psychiatry or occupational therapy.

[4]Notes on symptoms that exceed the sentence boundary occur in 3% of cases (4) *Sleep disorder* and (16) *Suicide/Self-extermination*, and in 1% of cases (11) *Alteration in efficiency/functionality*, (12) *Tiredness/Discouragement/Fatigue*, (13) *Despair*, (14) *Worry/Fear/Anxiety*, (18) *Sadness/Depressed mood* and (17) *Helplessness/Social harm/Loneliness*.

Symptom ID	Count	Tokens	Tokens annotation
1	15	235	15.7 (11.3)
2	16	138	8.6 (3.8)
3	17	158	9.3 (4.7)
4	31	301	9.7 (4.9)
5	31	329	10.6 (6.5)
6	35	288	8.2 (5.0)
7	40	397	9.9 (6.8)
8	43	436	10.1 (4.8)
9	73	1006	13.8 (8.2)
10	149	2104	14.1 (10.2)
11	155	1745	11.3 (6.5)
12	153	1576	10.3 (6.8)
13	164	2074	12.6 (8.2)
14	210	2523	12.0 (7.8)
15	234	2636	11.3 (8.4)
16	278	3103	11.2 (7.3)
17	377	5094	13.5 (9.1)
18	469	4580	9.8 (6.9)

Table 1: Amount of annotated instances for each symptom and its total number of tokens and average (and standard deviation) length number of tokens.

	Total	Avg (std)
Sentences	4711	7.80 (8.51)
Tokens	100013	165.58 (159.4)
Types	59106	97.86 (73.05)
Symptom	25112	41.58 (46.7)
Symptom (unique)	1568	2.60 (2.26)

Table 2: Corpus size in sentences, tokens, types, symptoms, and unique symptoms total and mean (and standard deviation) per text.

and processing difficulty, which are frequently associated with affective and cognitive states in mental health research.

We further analyzed the distribution of part-of-speech (POS) tags and dependency relations to detect potential shifts in syntactic organization and discourse structure[5]. Syntactic distributions provide a structural perspective on language production that may reflect changes in cognitive load, self-focus, or narrative framing.

In addition, we explored person usage, with particular attention to first-person singular forms. We measured the distribution of grammatical person (first, second, and third) in both singular and plural forms. The focus on person marking is motivated by evidence linking self-referential language to depressive symptomatology (De Choudhury et al., 2013; Nambisan et al., 2015; Ji et al., 2018).

Because Portuguese allows null subjects, we distinguished between explicit personal pronouns and subjects inferred from verbal morphology. This distinction enables a more accurate estimation of self-referential tendencies in a pro-drop language. We therefore computed counts both including and excluding explicit pronoun realizations.

Finally, we employed the Brazilian Portuguese adaptation of LIWC[6] (Filho et al., 2013) to quantify the use of words associated with affective and cognitive dimensions, including positive emotion, negative emotion, anxiety, anger, sadness, negation, certainty, death, cognitive processes, and general affect. LIWC categories were included to provide a psychologically grounded lexical profile aligned with established mental health research. We additionally examined the use of absolutist expressions[7], as such language has been associated with cognitive rigidity in depressive discourse.

To examine whether linguistic features differed as a function of depressive profile and symptom presence, statistical analyses were conducted at the subject level. Because most linguistic measures exhibited non-normal distributions, non-parametric tests were employed. Differences between PDP and non-PDP subjects were assessed using the Mann–Whitney U test. When comparing linguistic patterns across symptom categories, the Kruskal–Wallis test was applied, followed by Dunn's post-hoc pairwise comparisons when global significance was observed. To control for multiple comparisons across features and symptom groups, p-values were adjusted using the Benjamini–Hochberg False Discovery Rate (FDR) procedure. All analyses were conducted using subject-level normalized measures in order to control for variation in text length.

3.3 Symptoms and PDP linking

To investigate which symptoms are linked to possible depressive profiles (RQ3), we fitted a Bayesian Generalised Linear Mixed Model (GLMM) with a Bernoulli likelihood and logit link function. The dependent variable was the binary subject-level label (`is_pdp`). Predictors consisted of subject-level symptom counts derived from manual annotations.

Symptom counts were standardised (z-scores) to ensure comparability of regression coefficients. A varying intercept was included for each subject in order to account for subject-level heterogeneity and

[5]Syntactic analysis was performed using the *pt_core_news_lg* model in SpaCy.

[6]http://nilc.icmc.usp.br/portlex/index.php/pt/projetos/liwc

[7]List of absolutist expressions explored: tudo, todos, todas, sempre, nunca, ninguém, jamais, inteiramente, and completamente.

Sentence	Symptom	From a PDP post?
When I'm alone, I feel a deep emptiness inside myself.	Feeling of emptiness	PDP
I try to stay alive, just waiting for another day to pass.	Suicide/Self-extermination	PDP
My heart is beating so hard it hurts, and I just feel like trash and a total failure.	Physical symptom & Feeling of worthlessness/Low self-esteem	PDP
What do I do???	Difficulty in decision-making	PDP
People simply don't like me.	Helplessness/Social harm/Loneliness	PDP
If all this crap is well-being, then I don't know what BAD-BEING is.	Irritation/Aggressiveness	non-PDP
In these cases, I'm afraid of relapsing and doing things I promised myself I would never do again.	Worry/Fear/Anxiety	non-PDP
It hurts to see my friend depressed.	Sadness/Depressed mood	non-PDP
Parties and bars aren't as fun anymore; it's a completely empty world for me.	Loss/Diminishment of pleasure/libido & Feeling of emptiness	non-PDP

Table 3: Instances from the original corpus translated to English and slightly modified to anonymization purpose.

Linguistic characteristic	Level
Number of tokens	Lexical
Number of types	Lexical
Type-Token Ratio (TTR)	Lexical
Span length	Lexical
Mean syllables per word	Lexical
Flesch readability score	Lexical
Part-of-speech tags	Syntactic
Dependency relations	Syntactic
Grammatical person	Syntactic
LIWC categories	Psycholinguistic
Absolutist expressions	Psycholinguistic

Table 4: Linguistic characteristics investigated in this paper

Linguistic characteristic	PDP
Mean syllables per word	↓
Flesch readability score	↑
TTR	↓
POS tags	↑ VERB, ADV ↓ NOUN, ADJ, ADP, PROPN
Dependency relations	↑ advmod, obj, mark, xcomp ↓ amod, case, nmod, expl
Grammatical person	↑ 1_Sing VERB ↓ 3_Sing, 3_Plur VERB ↓ explicit subjects
LIWC categories	↑ negative emotion, anger, negation, certainty, death
Absolutist expressions	↑

Table 5: Linguistic characteristics that showed significant differences between PDP vs non-PDP posts. ↑ (↓) indicates a higher (lower) proportion of the linguistic phenomenon in PDP posts.

repeated symptom annotations within individuals. Formally, the model can be expressed as:

$$\text{logit}(P(\text{PDP}_i = 1)) = \alpha + \sum_{k=1}^{K} \beta_k X_{ik} + u_i, \quad (1)$$

where α is the global intercept, β_k are fixed effects for each symptom k, X_{ik} is the standardised count of symptom k for subject i, and $u_i \sim \mathcal{N}(0, \sigma_{\text{subject}})$ is the subject-level random intercept.

4 Results

4.1 PDP (RQ1)

The results for PDP versus non-PDP posts are described in this section. Table 5 summarizes the linguistic characteristics with significant differences in PDP posts regarding non-PDP ones.

Lexical. To investigate whether texts associated with possible depressive profiles (PDP) differ lexically from non-PDP texts, we compared several measures of lexical complexity.

No significant differences were observed in overall text length (tokens), vocabulary size (types), or mean sentence length (all p > .05). This suggests that both groups produce texts of comparable size and global lexical variety.

However, significant differences emerged in three readability-related measures. Texts in the PDP group exhibited a lower average number of syllables per word (p = .0002), higher Flesch readability scores (p = .0475) and lexical diversity (TTR; p = .0268), indicating simpler lexical choices and greater overall readability.

These results suggest that while lexical diversity remains stable across groups, texts associated with PDP tend to use shorter words and display lower lexical complexity. This pattern may reflect a more direct, less elaborated linguistic style.

Syntactic. The distribution of part-of-speech (PoS) categories revealed robust differences between PDP and non-PDP texts. Texts produced by individuals classified as PDP showed a significantly higher proportion of verbs (VERB) ($p < .001$) and adverbs (ADV) ($p < .001$), suggesting a more action- and modification-oriented discourse pro-

file. In contrast, PDP texts contained significantly fewer nouns (NOUN) ($p < .001$), adjectives (ADJ; $p = .005$), prepositions (ADP; $p = .0002$), proper nouns (PROPN; $p = .0001$), and subordinating conjunction (**SCONJ**; $p < .001$). These results indicate a structural shift from nominal and descriptive constructions toward verbal and adverbial constructions in PDP discourse. No significant differences were observed for determiners, auxiliaries, coordinating conjunctions, punctuation, numerals, or particles. Taken together, these findings suggest a reduction in nominal density and lexical specification in PDP texts, accompanied by increased predicate and modifier use.

PDP texts exhibited significantly higher proportions of adverbial modifiers (advmod) ($p < .001$), direct objects (obj) ($p = .0044$), subordinating markers (mark) ($p = .0007$), and open clausal complements (xcomp) ($p = .0422$). These increases suggest greater syntactic embedding and modification. Conversely, PDP texts showed significantly lower proportions of adjectival modifiers (amod) ($p < .001$), prepositional markers (case) ($p = .0001$), nominal modifiers (nmod) ($p < .001$), and expletives (expl) ($p = .0277$). The reduction in nominal modification and compounding further supports the interpretation of diminished nominal complexity in PDP discourse. Overall, the syntactic profile of PDP texts appears characterized by increased verbal projection and clausal embedding alongside reduced nominal modification.

PDP texts showed a significantly higher proportion of first-person singular verbs (1_Sing VERB) ($p < .001$), accompanied by significantly lower proportions of third-person singular (3_Sing VERB) ($p < .001$) and plural (3_Plur VERB) verbs ($p = .0002$). This pattern indicates a marked shift toward self-referential verbal constructions. Additionally, PDP texts contained significantly fewer explicit subjects ($p = .0292$). This suggests that increased self-focus may be structurally encoded through null subjects rather than overt pronouns.

Psycholinguistc. At the diagnostic level, several psycholinguistic dimensions significantly distinguished PDP from non-PDP posts. *Anger*-related lexicon was also elevated in PDP ($p = 0.033$), with means of 0.0172 (PDP) vs. 0.0133 (non-PDP). This suggests that irritability or anger-related expressions contribute to the linguistic differentiation of PDP. *Death*-related words strongly distinguished groups ($p < 0.001$). The PDP group (0.0146) used nearly twice as many death-related terms as the non-PDP group (0.0065), indicating marked lexical salience. *Affect*-related words were significantly elevated in PDP ($p = 0.0035$; 0.1036 vs 0.0959), reflecting a higher overall emotional lexical load. This pattern indicates that PDP narratives are characterized a generalized increase in affective language. PDP individuals produced significantly more *negative emotion* words ($p < 0.001$). Mean values were higher in the PDP group (0.0392) than in the non-PDP group (0.0327), indicating increased negative affective expression. *Certainty* words were more frequent in PDP ($p = 0.0043$; 0.0359 vs 0.0306). This may reflect more categorical or definitive cognitive positioning in PDP narratives. Indicating a small impact, *Absolutist* language was significantly higher in PDP ($p < 0.001$; 0.0084 vs. 0.0051), supporting previous literature linking cognitive rigidity and depressive symptomatology. *Negation* showed a robust difference ($p = 0.0001$). PDP individuals used more negation markers (0.0077) than non-PDP ones (0.0046), consistent with increased negative cognitive framing. *Sadness*-related words were also significantly more frequent in PDP ($p = 0.0055$; 0.0129 vs 0.0101). However, this category exhibited the smallest absolute difference between groups, suggesting that while sadness is statistically associated with PDP, it may represent a more pervasive baseline emotional marker rather than a strongly discriminative linguistic feature. Notably, *positive emotion* vocabulary did not reach statistical significance at the group level.

4.2 Symptom (RQ2)

In this section, we present the analysis to answer RQ2 on the symptoms of depression. Table 6 provides an overview of how the various linguistic features differ across each symptom. A systematic comparison of the differences between linguistic features can be found in Appendix A

Lexical. Our analysis of lexical variation across symptom annotations revealed significant differences across symptoms for token count, type count, TTR, and syllables per word. We observed that texts annotated with *Suicide/Self-extermination* (16) and *Physical symptom* (5) differed significantly from other symptoms in terms of the number of tokens and types and their ratio (TTR). Notably, *Physical symptom* (5) texts tended to be shorter and less lexically diverse, whereas *Suicide/Self-*

Symp	Lexical	Syntactic	Psycholinguistic
1	syllables ↑ (1)	–	–
2	–	1 Sing ↑ (1); 3 Sing ↓ (1)	–
3	syllables ↓ (1)	–	anxiety ↓ (1)
4	tokens ↑ (1); types ↑ (1); ttr ↓ (1)	NOUN ↑ (1)	anxiety ↓ (1)
5	tokens ↑ (6); types ↑ (6); ttr ↓ (2)	NOUN ↑ (1)	anxiety ↑ (2)
6	ttr ↓ (2); tokens ↑ (1)	NOUN ↓ (7); PRON ↑ (2); ADV ↑ (1)	positive ↑ (1); anxiety ↓ (1); death ↓ (1)
7	–	ADV ↑ (1); advmod ↑ (1)	–
8	tokens ↓ (1); types ↓ (1)	NOUN ↑ (1); 1 Sing ↑ (1); 3 Sing ↓ (1)	sadness ↑ (3); anxiety ↓ (1)
9	–	ADV ↑ (1); NOUN ↓ (1); PROPN ↓ (1)	anxiety ↓ (1); death ↓ (1)
10	tokens ↑ (2); types ↑ (2); ttr ↓ (1)	1 Sing ↓ (12); ADV ↓ (10); advmod ↓ (10)	positive ↓ (3); anxiety ↓ (2); sadness ↓ (2)
11	tokens ↓ (1); types ↓ (1)	PRON ↓ (4); NOUN ↑ (2); ADV ↑ (1)	anxiety ↓ (1); death ↓ (1)
12	tokens ↑ (1); types ↑ (1)	ADV ↑ (1); PROPN ↓ (1); advmod ↑ (1)	sadness ↑ (3); anxiety ↓ (1); death ↓ (1)
13	tokens ↑ (1); types ↑ (1)	advmod ↑ (2); ADV ↑ (1); NOUN ↓ (1)	anxiety ↓ (1); death ↓ (1)
14	tokens ↓ (1); types ↓ (1)	NOUN ↑ (6); PRON ↓ (2); ADP ↑ (1)	anxiety ↑ (13); sadness ↓ (2); death ↓ (1)
15	tokens ↑ (1); types ↑ (1)	NOUN ↓ (4); PRON ↑ (3); case ↓ (2)	affect total ↑ (1); positive ↑ (1); anxiety ↓ (1)
16	tokens ↓ (8); types ↓ (7); ttr ↑ (4)	ADV ↑ (1); NOUN ↓ (1); PRON ↑ (1)	death ↑ (10); anxiety ↓ (2); sadness ↓ (2)
17	tokens ↑ (1); tokens ↓ (1); types ↑ (1)	NOUN ↓ (2); ADV ↑ (1); PRON ↑ (1)	positive ↑ (1); anxiety ↓ (1); death ↓ (1)
18	tokens ↓ (2); types ↓ (2); ttr ↑ (2)	NOUN ↑ (2); ADV ↑ (1); advmod ↑ (1)	anxiety ↓ (1); death ↓ (1)

Table 6: Top 3 dominant feature-direction combinations per symptom. The numbers in brackets indicate the number of symptoms that are statistically different.

extermination (16) texts were comparatively longer and less lexically concentrated. The difference in the number of syllables appears only between *Attention/memory deficit* (2) and *Alteration in weight/eating habits* (3), with 3 tending to have words with slightly more syllables (1.77 vs 1.73).

Syntactic. The symptom *Irritation/Aggressiveness* (10) systematically has less adverbial usage (PoS ADV: and DEP advmod) compared with other symptoms. This might suggest that a trigger for identifying symptom 10 is the verb modification. We also observe a difference in the proportion of noun usage in some symptoms. The *Difficulty in decision-making* (6) showed the highest proportion of nouns, while *Worry/Fear/Anxiety* (14) and *Irritation/Aggressiveness* (10) had the smallest proportion. These findings indicate that symptom clusters differ in the degree of nominal elaboration and referential density. Proper noun usage differed significantly across symptoms ($p = .0001$), with repeated contrasts involving *Irritation/Aggressiveness* (10). This may reflect differences in narrative framing. This might indicate a tendency to indicate references to named entities, people or institutions targeted.

Looking in more detail, targeting the pronoun usage, we also observe a reduction in the use of *first-person singular* in the *Irritation/Aggressiveness* (10) symptom.

Psycholinguistc. At the symptom level, significant differences emerged for several affective and cognitive dimensions.

A global difference in *affect* was observed across symptoms ($p = 0.0062$). Pairwise comparisons indicated that *Irritation/Aggressiveness* (10) differed significantly from *Feeling of worthlessness/Low self-esteem* (15). Inspection of group means suggests that 10 presented one of the highest affective loads, whereas 15 showed comparatively lower values. This indicates heterogeneity in overall affective expression across symptom categories.

Positive emotion varied significantly by symptom. Differences primarily involved *Irritation/Aggressiveness* (10), which showed higher positive emotion levels. Overall, decision-related narratives exhibited the highest positive emotion proportion, whereas irritability-related narratives showed among the lowest.

Anxiety showed the strongest symptom-level effect ($p < 0.0001$). As expected, *Worry/Fear/Anxiety* (14) presented the highest mean anxiety proportion (0.0168), significantly exceeding multiple other symptom categories. This confirms that anxiety-related symptomatology is linguistically reflected in increased anxiety lexicon. To illustrate that Figure 1 show the word cloud for the text spans annotated with *Worry/Fear/Anxiety* (211 instances), where we can see the highlighted words *medo* (fear), with 76 occurrences, and *ansiedade* (anxiety), with 74 occurrences.

Sadness also differed across symptoms ($p = 0.0005$). Smallest sadness proportions were observed in *Tiredness/Discouragement/Fatigue*, 12, (0.016). This suggests that symptoms related to fatigue and lack of enjoyment are more strongly associated with the explicit lexicon of sadness.

Death-related words showed the clearest differentiation ($p < 0.001$). *Suicide/Self-extermination* (16) exhibited substantially higher death lexicon proportions (0.0214) than all other symptom cate-

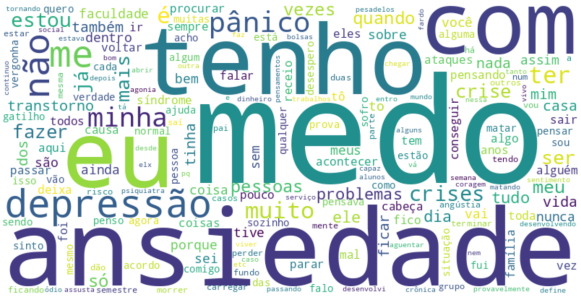


Figure 1: Word cloud for the text spans annotated with *Worry/Fear/Anxiety*

gories. This pattern is theoretically coherent and supports the construct validity of the symptom annotation.

4.3 Symptoms and PDP (RQ1 & RQ2)

Now that we identified the link between the linguistic marker related to the PDP and the symptoms, we explore a possible interplay between RQ1 and RQ2. In other words, can the linguistic characteristics of symptoms differ between PDP and non-PDP?

Lexical. Overall, lexical complexity differences between PDP and non-PDP are not uniform across symptoms. However, symptoms such as *suicide, despair, and low self-esteem* show a consistent pattern of structural compression and increased readability in PDP contexts. This suggests that depressive profile status does not merely increase symptom frequency but is associated with subtle shifts in linguistic organization.

Syntactic. The analysis revealed a robust and structured pattern of syntactic differentiation across several grammatical domains. Regarding POS distribution, significant interaction effects were observed.[8] Crucially, these effects were not uniform but driven by specific symptom–PDP combinations. For instance, *verb* usage differed significantly between *Suicide/Self-extermination-PDP* and *Helplessness/Social harm/Loneliness-non-PDP* ($p = 0.0239$), as well as between *Feeling of worthlessness/Low self-esteem-PDP* and *Helplessness/Social harm/Loneliness-non-PDP* ($p = 0.0208$). *Adverbial* usage showed a consistent pattern: *Irritation/Aggressiveness-non-PDP* differed from multiple PDP-marked symptom groups (e.g., *Sadness/Depressed mood*, *Despair* and *Feeling of guilt* in PDPs; $p \leq 0.0389$), suggesting that adverbial modulation may differentiate emotionally reactive non-PDP irritation from more internally focused symptom constellations. *Noun* and *preposition* distributions further distinguished PDP-marked low self-worth (*Feeling of worthlessness/Low self-esteem-PDP*) from anxiety-related non-PDP cases (*Worry/Fear/Anxiety-non-PDP*), indicating structural shifts in nominal framing and relational marking.

Dependency-level analyses reinforced these findings.[9] The *adverbial modifier* pattern mirrored the POS-level *adverb* effect, with *Irritation/Aggressiveness-non-PDP* contrasting systematically with multiple PDP symptom groups. Additionally, *case* and *mark* dependencies showed differences primarily between *Feeling of worthlessness/Low self-esteem-PDP* and *Worry/Fear/Anxiety-non-PDP*, suggesting alterations in clausal embedding and argument structure associated with depressive self-referential symptoms. These results indicate that depressive profiles are not characterized solely by lexical shifts but by reorganizations in clause architecture and modifier integration.

Verb-person morphology exhibited one of the strongest interaction effects. *First-person singular* verbs showed significant differences ($p < 0.0001$), with *Irritation/Aggressiveness-non-PDP* differing from nearly all PDP-marked symptom groups.

Similarly, *third-person singular* verbs ($p < 0.0001$) distinguished fatigue- and apathy-related PDP groups from several non-PDP symptom categories. These patterns suggest that symptom configurations interact with PDP status in shaping perspectival anchoring: PDP contexts tend to intensify first-person singular marking, while non-PDP irritation shows comparatively distinct distributional behavior. Pronoun-level interactions were more selective. *Third-person singular pronouns* ($p = 0.0006$) differentiated several PDP-marked symptoms (e.g., *Tiredness/Discouragement/Fatigue-PDP*, *Suicide/Self-extermination-PDP*) from *Sadness/Depressed mood-non-PDP*. *First-person singular pronouns* ($p = 0.0005$) also exhibited interaction effects, reinforcing the verb-based

[8] We found significant interaction effects for verbs ($p = 0.001$), adverbs ($p = 0.001$), nouns ($p < 0.0001$), prepositions ($p = 0.0083$), subordinating conjunction ($p = 0.0326$), pronouns ($p = 0.0011$), and proper nouns ($p < 0.0001$).

[9] Significant interaction effects emerged for *adverbial modifier* ($p = 0.0001$), *case* ($p = 0.0027$), *nominal dependents* ($p = 0.0231$), *mark* ($p = 0.0118$), *auxiliary* ($p = 0.0124$), and *indirect object* ($p = 0.0136$).

self-referential pattern. Oblique pronouns showed no significant interaction effects, indicating that argument-level object marking remains relatively stable across symptom–PDP combinations.

Overall, the interaction results demonstrate that depressive linguistic profiles are not reducible to either symptom type or PDP status alone. Rather, specific symptom configurations acquire distinct syntactic signatures when co-occurring with PDP. The most consistent interaction markers involve (i) verbal density and person marking, (ii) adverbial modification, and (iii) nominal and relational structuring. These findings support the hypothesis that PDP amplifies or reshapes the grammatical realization of certain symptom, particularly self-reference, cognitive evaluation and affective modulation.

Psycholinguistc. The clearest descriptive amplification of PDP effects occurs within *Suicide/Self-extermination* (16). PDP participants in this category show markedly higher *death* lexicon proportion (0.0223) compared to non-PDP individuals with the same symptom (0.0100). Similar, though smaller, PDP-related increases appear in *Feeling of worthlessness/Low self-esteem* (15) and *Sadness/Depressed mood* (18).

Within *Feeling of guilt* (9), PDP individuals show higher *negative emotion* (0.0477) compared to non-PDP (0.0346). A similar pattern appears in *Sadness/Depressed mood* (18) and *Suicide/Self-extermination* (16), suggesting that PDP intensifies negative affective expression.

PDP participants generally show higher *negation* and *absolutist* usage within most symptom categories (e.g., *Feeling of worthlessness/Low self-esteem* (15), *Loss/Diminishment of pleasure/libido* (8)), suggesting that diagnostic status amplifies cognitively rigid and negatively framed language beyond symptom effects alone.

4.4 Effects of symptoms on PDP (RQ3)

All chains converged satisfactorily ($\hat{R} \approx 1.00$ for all parameters). The subject-level variance parameter ($\sigma_{\text{subject}} = 0.339$) indicates moderate heterogeneity across individuals.

Table 7 summarizes posterior estimates for symptom coefficients (β), their 94% Highest Density Intervals (HDI), and the corresponding odds ratios (OR = e^{β}).

Suicidal ideation (16) exhibited the strongest association with PDP ($\beta = 1.516$), corresponding to an OR of approximately 4.55. This indicates that a one-standard-deviation increase in this symptom multiplies the odds of PDP by four. *Despair* (13) also showed a strong positive effect (OR ≈ 2.30), doubling the odds of PDP. Core affective symptoms such as sadness (18), despair (15) and functional impairment (11) nearly doubled the odds as well.

Irritation/Aggressiveness (10) showed a negative association ($\beta = -0.511$), with an OR of 0.60. This suggests that, controlling for other symptoms, higher levels of irritability were associated with lower odds of PDP in this dataset.

Symptom ID	β	94% HDI	OR
16	1.516	[1.064; 2.021]	4.55
13	0.835	[0.385; 1.313]	2.30
18	0.762	[0.466; 1.037]	2.14
15	0.579	[0.185; 0.959]	1.78
11	0.568	[0.248; 0.908]	1.76
8	0.439	[0.104; 0.805]	1.55
17	0.437	[0.154; 0.719]	1.55
7	0.435	[-0.037; 0.953]	1.54
1	0.375	[-0.011; 0.786]	1.45
2	0.333	[-0.036; 0.755]	1.40
12	0.171	[-0.156; 0.509]	1.19
5	0.106	[-0.132; 0.374]	1.11
9	0.098	[-0.188; 0.407]	1.10
14	-0.015	[-0.259; 0.225]	0.99
4	-0.111	[-0.463; 0.225]	0.89
3	-0.115	[-0.406; 0.142]	0.89
6	-0.210	[-0.497; 0.079]	0.81
10	-0.511	[-0.772; -0.228]	0.60

Table 7: Posterior estimates for symptom effects on PPD.

5 Conclusions

This paper focused on characterising the language of possible depressive profile narratives, looking at lexical, syntactic and psycholinguistic clues in texts and in the symptoms in them. For RQ1 and RQ2 the results confirm that there are relevant signals, but more investigation is needed for their links with PDP and their interplay (§ 4.1 and 4.2). For RQ3, we found symptoms closely associated with PDP (§ 4.4). These findings shed light on how symptoms of depression appear on social media, opening up new avenues in a range of fields, such as clinical practice and the development of more accurate computational tools, including LLMs that can warn of potential cases where immediate intervention is needed. Future work includes extending the characterisation to include analyses of the incidence of figurative language and also replicate this analysis on English language since AIM-Health project is an international research cooperation between UK and Brazil.

Ethical Considerations

The data for this research comprises anonymous text posts automatically collected from a public social media platform. These posts are part of a corpus made available to researchers upon request at no cost. No information that could identify the authors of the posts is present in the texts, thereby protecting their privacy and minimizing potential harm. Researchers were informed that the posts' content may include sensitive material related to mental health. All analyses were conducted using established, widely recognized methods.

Limitations

This analysis reflects the contents of the corpus and any biases derived from the corpus construction regarding the social network (Facebook) selection, the keywords and protocols applied. The annotation may also be biased, even though it was carried out by psychology, psychiatry and occupational therapy students after training with experienced mental health professionals and with their review. However, it is important to emphasize that for the purpose of the work, a false positive is better than a false negative; that is, it is better to find an indication of depression that is not exactly depression and invite the person to talk than to miss an indication of depression in someone who really should be invited for a conversation. Another limitation of this work is the small corpus size. However, although small, the collection contains rich data that should be considered representative of the problem: depression among university students.

Acknowledgments

This study was financed, in part, by the São Paulo Research Foundation (FAPESP), Brasil. Process Number #2024/10233-7 (AIM-Health project). This project is also financed by UKRI/MRC – UK Research and Innovation / Medical Research Council. We thank the Amive project team for giving us access to the corpus. The authors thank the support received from the World University Network (WUN) RDF.

References

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3(null):993–1022.

Margaret M Bradley and Peter J Lang. 1999. Affective norms for english words (anew): Instruction manual and affective ratings. Technical report, Technical report C-1, The center for research in psychophysiology, University of Florida.

Vinicius Casani, Alinne C. Correa Souza, Rafael G. Mantovani, and Francisco Carlos M. Souza. 2021. DP-symptom-identifier: uma estratégia para classificar sintomas de depressão utilizando um conjunto de dados textuais na língua portuguesa. In *Annals of the XIII Brazilian Symposium of Information Tecnology and Human Language (STIL 2021)*. SBC.

Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013. Predicting depression via social media. In *Proceedings of the international AAAI conference on web and social media*, volume 7, pages 128–137.

Antonia Errazuriza, Dalia Avello-Vegab, Juan P. Ramirez-Mahalufa, Rafael Torresa, Nicolas A. Crossleya, Eduardo A. Undurragac, and Peter B. Jones. 2023. Prevalence of depressive disorder in the adult population of latin america: a systematic review and meta-analysis. *The Lancet Regional Health – Americas*, 26(100587).

Pedro P. Balage Filho, Thiago Alexandre Salgueiro Pardo, and Sandra M. Aluísio. 2013. An evaluation of the Brazilian Portuguese LIWC dictionary for sentiment analysis. In *Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology*.

Global Burden of Disease Collaborative Network. 2025. *Global Burden of Disease Study 2023 (GBD 2023) Covariates 1980-2023*. Institute for Health Metrics and Evaluation (IHME), Seattle, United States of America.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Shaoxiong Ji, Celina Ping Yu, Sai-fu Fung, Shirui Pan, Guodong Long, and Gao Cong. 2018. Supervised learning for suicidal ideation detection in online user content. *Complex.*, 2018.

Paulo Mann, Aline Paes, and Elton H. Matsushima. 2020. See and read: Detecting depression symptoms in higher education students using multimodal social media data. *Proceedings of the International AAAI Conference on Web and Social Media*, 14:440–451.

Paulo Mann, Matheus Yasuo Ribeiro Utino, Elton Hiroshi Matsushima, and Aline Paes. 2025. The topics of depression on social networking sites. *Journal of the Brazilian Computer Society*, 31(1):771–806.

Teresa BF Martins, Claudete M Ghiraldelo, Maria das Graças Volpe Nunes, and Osvaldo Novais de Oliveira Junior. 1996. Readability formulas applied to textbooks in brazilian portuguese.

Augusto R. Mendes and Helena Caseli. 2024. Identifying fine-grained depression signs in social media posts. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8594–8604, Torino, Italia. ELRA and ICCL.

Priya Nambisan, Zhihui Luo, Akshat Kapoor, Timothy B. Patrick, and Ron A. Cisler. 2015. Social media, big data, and public health informatics: Ruminating behavior of depression revealed through twitter. In *2015 48th Hawaii International Conference on System Sciences*, pages 2906–2913.

J.W. Pennebaker, R.L. Boyd, K. Jordan, and K. Blackburn. 2015. The development and psychometric properties of LIWC2015. Austin, TX: University of Texas at Austin.

Wesley Ramos dos Santos, Rafael Lage de Oliveira, and Ivandré Paraboni. 2023. SetembroBR: a social media corpus for depression and anxiety disorder prediction. *Language Resources and Evaluation*, pages 1–28.

Yla R Tausczik and James W Pennebaker. 2010. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social psychology*, 29(1):24–54.

Raluca Nicoleta Trifu, Bogdan Nemeș, Dana Cristina Herta, Carolina Bodea-Hategan, Dorina Anca Talaș, and Horia Coman. 2024. Linguistic markers for major depressive disorder: a cross-sectional study using an automated procedure. *Frontiers in Psychology*, Volume 15 - 2024.

World Health Organization. 2021. Comprehensive mental health action plan 2013–2030. Technical report, World Health Organization.

Shweta Yadav, Jainish Chauhan, Joy Prakash Sain, Krishnaprasad Thirunarayan, Amit Sheth, and Jeremiah Schumm. 2020. Identifying depressive symptoms from tweets: Figurative language enabled multitask learning framework. *arXiv preprint arXiv:2011.06149*.

Amir Hossein Yazdavar, Hussein S Al-Olimat, Monireh Ebrahimi, Goonmeet Bajaj, Tanvi Banerjee, Krishnaprasad Thirunarayan, Jyotishman Pathak, and Amit Sheth. 2017. Semi-supervised approach to monitoring clinical depressive symptoms in social media. In *Proceedings of the 2017 IEEE/ACM international conference on advances in social networks analysis and mining 2017*, pages 1191–1198.

A Symptom-Level Pairwise Linguistic Contrasts

A.1 Lexical contrasts

A.2 Syntactic contrasts

A.3 Psycholinguistic contrasts

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1	–		syllables↑															
3	syllables↓		–															
4				–												tk↑, ttr↓, ty↑		
5					–			tk↑, ty↑			tk↑, ty↑			tk↑, ty↑		tk↑, ttr↓, ty↑	tk↑, ty↑	tk↑, ttr↓, ty↑
6						–										tk↑, ttr↓		ttr↓
8					tk↓, ty↓		–											
10								–								tk↑, ttr↓, ty↑		tk↑, ty↑
11					tk↓, ty↓					–								
12											–					tk↑, ty↑		
13												–				tk↑, ty↑		
14					tk↓, ty↓								–					
15				tk↓, ttr↑, ty↓	tk↓, ttr↑, ty↓	tk↓, ttr↑				tk↓, ttr↑, ty↓		tk↓, ty↓			–	tk↑, ty↑	tk↓, ty↓	
16					tk↓, ty↓										tk↓, ty↓	–		
17					tk↓, ttr↑, ty↓	ttr↑				tk↓, ty↓						tk↑, ty↑	–	
18																		–

Table 8: Lexical symptom contrasts. Cells indicate significant differences; arrows denote direction (row relative to column).

	2	4	5	6	7	8	9	10
2	–							1 Sing↑, 3 Sing↓
4		–		NOUN↑				
5			–	NOUN↑				
6		NOUN↓	NOUN↓	–		NOUN↓		1 Sing↑, ADV↑, NOUN↓, advmod↑
7					–			ADV↑, advmod↑
8				NOUN↑		–		1 Sing↑, 3 Sing↓
9							–	1 Sing↑, 3 Sing↓, ADV↑, PROPN↓, advmod↑
10	1 Sing↓, 3 Sing↑			1 Sing↓, ADV↓, NOUN↑, advmod↓	ADV↓, advmod↓	1 Sing↓, 3 Sing↑	1 Sing↓, 3 Sing↑, ADV↓, PROPN↑, advmod↓	–
11				NOUN↑, PRON↓				1 Sing↑, 3 Sing↓, ADV↑, advmod↑
12								1 Sing↑, ADV↑, PROPN↓, advmod↑
13								1 Sing↑, 3 Sing↓, ADV↑, PROPN↓, advmod↑
14				NOUN↑, PRON↓			NOUN↑	1 Sing↑
15								1 Sing↑, 3 Sing↓, ADV↑, NOUN↓, PRON↑, PROPN↓, advmod↑, case↓
16								1 Sing↑, 3 Sing↓, ADV↑, PROPN↓, advmod↑
17								1 Sing↑, 3 Sing↓, ADV↑, NOUN↓, PROPN↓, advmod↑
18				NOUN↑				1 Sing↑, ADV↑, advmod↑

Table 9: Syntactic symptom contrasts (from symptoms 1 to 10). Cells indicate significant differences; arrows denote direction (row relative to column).

	11	12	13	14	15	16	17	18
2								
4								
5								
6	NOUN↓, PRON↑			NOUN↓, PRON↑				NOUN↓
7								
8								
9				NOUN↓				
10	1 Sing↓, 3 Sing↑, ADV↓, advmod↓	1 Sing↓, ADV↓, PROPN↑, advmod↓	1 Sing↓, 3 Sing↑, ADV↓, PROPN↑, advmod↓	1 Sing↓	1 Sing↓, 3 Sing↑, ADV↓, NOUN↑, PRON↓, PROPN↑, advmod↓, case↑	1 Sing↓, 3 Sing↑, ADV↓, PROPN↑, advmod↓	1 Sing↓, 3 Sing↑, ADV↓, NOUN↑, PROPN↑, advmod↓	1 Sing↓, ADV↓, advmod↓
11	–				NOUN↑, PRON↓	PRON↓	PRON↓	
12		–						
13			–	NOUN↓, advmod↑				
14			NOUN↑, advmod↓	–	ADP↑, NOUN↑, PRON↓, case↑	NOUN↑	NOUN↑	
15	NOUN↓, PRON↑			ADP↓, NOUN↓, PRON↑, case↓	–			NOUN↓
16	PRON↑			NOUN↓		–		
17	PRON↑			NOUN↓			–	
18					NOUN↑			–

Table 10: Syntactic symptom contrasts (from symptoms 11 to 18). Cells indicate significant differences; arrows denote direction (row relative to column).

	3	4	5	6	8	9	10	11	12	13	14	15	16	17	18
3	–										anxiety↓				
4		–									anxiety↓				
5			–				anxiety↑						anxiety↑		
6				–			positive↑				anxiety↓		death↓		
8					–		sadness↑				anxiety↓, sadness↑		sadness↑		
9						–					anxiety↓		death↓		
10			anxiety↓	positive↓	sadness↓		–		sadness↓		anxiety↓	affect↓, positive↓	death↓	positive↓	
11								–			anxiety↓		death↓		
12							sadness↑		–		anxiety↓, sadness↑		death↓, sadness↑		
13										–	anxiety↓		death↓		
14	anxiety↑	anxiety↑		anxiety↑	anxiety↑, sadness↓	anxiety↑	anxiety↑	anxiety↑	anxiety↑, sadness↓	anxiety↑	–	anxiety↑	anxiety↑, death↓	anxiety↑	anxiety↑
15							affect↑, positive↑				anxiety↓	–	death↓		
16			anxiety↓	death↑	sadness↓	death↑	death↑	death↑	death↑, sadness↓	death↑	anxiety↓, death↑	death↑	–	death↑	death↑
17							positive↑				anxiety↓		death↓	–	
18											anxiety↓		death↓		–

Table 11: Psycholinguistic symptom contrasts. Cells indicate significant differences; arrows denote direction (row relative to column).

Caracterização lexical e sintática de notícias falsas em português produzidas por humanos e por máquinas

Pedro Lucas Castro de Andrade, Renato Moraes Silva, Thiago Alexandre Salgueiro Pardo
Núcleo Interinstitucional de Inteligência Computacional (NILC)
Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo
São Carlos – SP – Brasil
pedroandrade@usp.br, {renatoms,taspardo}@icmc.usp.br

Resumo

Notícias falsas são um grande problema para a sociedade. Com a Inteligência Artificial generativa, notícias falsas produzidas pela máquina têm se proliferado, tornando o cenário mais desafiador. Apesar da relevância desse problema, em línguas sub-representadas como o Português, as pesquisas que buscam diferenciar notícias falsas de humanos e de máquinas são incipientes. Buscando preencher essa lacuna, este artigo explora os corpora Fake.br e FakeTrueBR expandidos com notícias falsas geradas automaticamente, caracterizando lexical e sintaticamente as notícias falsas produzidas por humanos e por máquina. Os resultados mostram que textos gerados por máquina apresentam palavras significativamente mais longas, maior uso de modificadores adjetivais e menor diversidade sintática, apesar de utilizarem mais regras sintáticas por sentença. Em contrapartida, textos humanos exibem maior variabilidade estilística em todas as dimensões analisadas.

1 Introdução

Notícias falsas são um grande mal na sociedade moderna, especialmente em função de sua facilidade de disseminação pelas redes sociais. Mais recentemente, com a popularização da Inteligência Artificial generativa, o desafio se tornou ainda maior, dada a facilidade de produção desse tipo de conteúdo por máquinas, com estilo de escrita similar ao dos humanos (Su et al., 2024a; Chen e Shu, 2024).

Silva et al. (2024) compilam diversas iniciativas de detecção automática de notícias falsas para o português, sendo que o Fake.Br (Monteiro et al., 2018; Silva et al., 2020) foi o primeiro corpus de notícias falsas criado para esta língua e o primeiro a ser testado em métodos de detecção. Desde então, diversos esforços surgiram no enfrentamento do problema das notícias falsas.

Apesar dos avanços, pelo que se tem conhecimento, não há iniciativas para o português de tentar diferenciar automaticamente notícias falsas produzidas por humanos das produzidas por máquina. De fato, apenas recentemente o trabalho de Silva et al. (2025) disponibilizou corpora de notícias falsas produzidas por máquina para o português. Para criá-las, os autores utilizaram como base os corpora Fake.br (Monteiro et al., 2018; Silva et al., 2020) e FakeTrueBR (Chavarro et al., 2023), que contêm pares de notícias alinhadas, isto é, para cada notícia falsa escrita por humano, há uma notícia verdadeira relacionada, coletada de agências jornalísticas oficiais. O estudo conduzido por Silva et al. (2025) forneceu essas notícias verdadeiras como entrada para o Sabiá-3 (Pires et al., 2023; Abonizio et al., 2024), um grande modelo de língua (LLM, do inglês *Large Language Model*), solicitando, via *prompt*, a produção de versões falsas. No total, geraram 3.600 notícias falsas para o Fake.br e 1.791 para o FakeTrueBR, igualando o número de notícias falsas produzidas por humanos nesses corpora.

Neste artigo, realiza-se a análise dessas notícias, caracterizando lexical e sintaticamente as notícias falsas produzidas por humanos e por máquina, buscando atributos discriminativos desses textos. Em particular: levanta-se o vocabulário utilizado nessas notícias, avaliando se há diferenças nos termos mais utilizados; analisa-se a complexidade silábica dos textos, comparando o tamanho médio das palavras e sentenças; verifica-se a distribuição de classes gramaticais; extraem-se e se comparam regras sintáticas de formação de sentenças.

Esse esforço é especialmente motivado pela necessidade de identificar a origem do conteúdo falso em uma época em que a Inteligência Artificial generativa tem dominado as discussões científicas e na sociedade em geral. Compreender as diferenças linguísticas entre textos produzidos por humanos e por máquinas é fundamental para o desenvolvimento de sistemas de detecção mais robustos.

Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)
Vol. 2, pages 148–158
April 13-16, 2026

O restante deste artigo está organizado da seguinte forma: a Seção 2 apresenta os trabalhos relacionados; a Seção 3 descreve a caracterização das notícias falsas; na Seção 4, discutem-se os achados e apresentam-se algumas considerações finais.

2 Trabalhos Relacionados

A detecção de notícias falsas e a identificação de textos gerados por LLMs constituem áreas de pesquisa que, embora distintas, têm se aproximado significativamente nos últimos anos. Esta seção apresenta brevemente os principais trabalhos nessas frentes que fundamentam esta pesquisa.

As abordagens tradicionais para detecção de notícias falsas concentram-se em textos humanos e exploram diferentes aspectos linguísticos e contextuais. Rashkin et al. (2017) e Pérez-Rosas et al. (2018) investigaram padrões linguísticos e estilos de escrita como indicadores de falsidade. Horne e Adalı (2017) identificaram que notícias falsas tendem a utilizar vocabulário mais emocional, textos mais simples e tom sensacionalista. Outras abordagens incluem análise semântica e estrutural do conteúdo (Jin et al., 2017; Zhou e Zafarani, 2020), verificação automática de fatos (Graves e Cherubini, 2016), análise de credibilidade da fonte (Baly et al., 2020) e uso de contexto social de redes (Barnabò et al., 2022). No contexto da língua portuguesa, Silva et al. (2024) apresentam uma visão abrangente dos métodos de detecção, incluindo abordagens baseadas em aprendizado de máquina com atributos linguísticos, modelos de linguagem pré-treinados e técnicas de verificação de conteúdo.

Os métodos de detecção de textos produzidos por máquinas (independentemente da veracidade dos fatos) podem ser categorizados em diferentes abordagens. Métodos baseados em características linguísticas exploram diferenças estatísticas entre textos humanos e gerados por máquina. Shah et al. (2023) demonstraram que características estilísticas como contagem de sílabas, comprimento de palavras e estrutura de sentenças permitem distinguir textos com precisão de 93%. Classificadores baseados em características também têm sido amplamente utilizados, alcançando resultados expressivos por meio da extração de atributos textuais combinados com técnicas de aprendizado de máquina (Aich et al., 2022).

Métodos baseados em modelos neurais, especialmente *transformers*, têm apresentado os melhores resultados em tarefas de detecção. Modelos como BERT (Devlin et al., 2019) e suas variantes são frequentemente usados para classificação, alcançando taxas de precisão superiores a 95% em cenários intra-domínio e superando métodos *zero-shot* como DetectGPT (Mitchell et al., 2023), apesar de degradarem em domínios não vistos durante o treinamento (Su et al., 2024b). Deve-se destacar a existência de competições como a AuTexTification, que, em 2024, incluiu o português, sendo que o sistema vencedor utilizou um *ensemble* de *transformers* multilíngues (DistilBERT, mDeBERTa-v3, XLM-RoBERTa) com regressão logística, alcançando Macro-F1 de 0.805 (Fernández García e Segura-Bedmar, 2024).

Na interseção dessas linhas de pesquisa, Su et al. (2023) demonstraram que detectores de notícias falsas apresentam viés contra textos gerados por LLMs, o que motiva investigações mais profundas. Su et al. (2024b) desvinculam a veracidade da origem do texto, revelando que detectores treinados exclusivamente com textos humanos generalizam melhor para textos gerados por máquina do que o inverso. Zhou et al. (2023) identificaram que textos de IA em redes sociais tendem a ser menos autênticos, mais emocionais e menos analíticos, enquanto textos noticiosos de IA são mais formais e analíticos. Além disso, a IA utiliza menos expressões informais e gírias, mas emprega mais palavras emocionais e expressões de raciocínio. Berber Sardinha (2024) avaliou múltiplos domínios utilizando análise multidimensional e observou que, para notícias, há diferenças sistemáticas em dimensões linguísticas. Para o português, Silva et al. (2025) geraram notícias falsas sintéticas utilizando o modelo Sabiá-3 (Abonizio et al., 2024; Pires et al., 2023), um LLM específico para português, e avaliaram o impacto em classificadores de aprendizado de máquina. Seus resultados indicam que a natureza dinâmica das notícias falsas e a escassez de recursos para idiomas além do inglês representam desafios significativos para a área.

Inspirado nesses trabalhos, este artigo foca na caracterização linguística comparativa entre notícias falsas escritas por humanos e geradas por LLM. São empregadas principalmente análises lexicais e sintáticas detalhadas para identificar traços discriminativos dos dois grupos, visando fornecer uma análise linguística aprofundada e inédita do fenômeno para o português, assim como subsidiar métodos mais robustos de detecção de notícias falsas no futuro.

3 Caracterização das Notícias Falsas

Este trabalho usa os únicos corpora disponíveis para o português que contam com notícias falsas geradas por humanos e por máquina: o Fake.br (Monteiro et al., 2018; Silva et al., 2020) e o FakeTrueBR (Chavarro et al., 2023), cuja geração de notícias falsas por máquina foi conduzida por Silva et al. (2025).

Em particular, os dados do Fake.br foram pré-processados, fazendo-se adequações de formato, como correção de codificação de símbolos especiais e ajuste na segmentação sentencial e na tokenização, deixando os corpora em igualdade de condições para análise e comparação. Isso foi necessário em função das origens diferentes dos corpora e suas decisões de coleta e armazenamento.

A Tabela 1 apresenta as características básicas dos corpora, já separados pela origem das notícias falsas[1]. Nota-se que há variações importantes: as notícias do corpus FakeTrueBR são menores do que as do Fake.br, e as notícias produzidas por máquina são consideravelmente mais extensas do que as humanas em ambos os corpora (368,78 contra 191,14 *tokens* por notícia no Fake.br). Além disso, as sentenças geradas por máquina tendem a ser mais longas (20,57 contra 15,58 *tokens* por sentença no Fake.br). Essas diferenças são levadas em consideração nas análises feitas a seguir.

O Quadro 1 mostra notícias falsas do corpus Fake.br produzidas por humano e por máquina. O texto humano apresenta linguagem informal e emotiva, com uso de exclamações (*"Vergonha!"*), gírias (*"calote"*) e opiniões explícitas do autor (*"roubarem descaradamente"*). Em contraste, o texto gerado por máquina adota um tom formal, com estruturas sintáticas mais elaboradas, uso de fontes anônimas (*"Segundo fontes exclusivas"*) e adjetivos intensificadores (*"suposta dívida milionária"*). Ademais, o texto de máquina é mais extenso e apresenta maior coesão textual, enquanto o texto humano é mais fragmentado e direto.

3.1 Vocabulário

Conduziu-se um levantamento das palavras mais frequentes nas notícias falsas produzidas por humanos e por máquina. A análise vocabular é motivada pelo resultados de Horne e Adalı (2017), que mostraram que textos falsos humanos tendem a utilizar vocabulário mais emocional e sensacionalista, enquanto Zhou et al. (2023) observaram que textos noticiosos gerados por IA empregam menos expressões informais e mais palavras de raciocínio. Essas evidências sugerem que diferenças vocabulares podem servir como indicadores da origem humana ou de máquina das notícias falsas.

Para tanto, a fim de mitigar ruídos, todas as palavras foram normalizadas para letra minúscula. Na Tabela 2, são mostradas as dez palavras mais frequentes nos corpora (ordenadas das mais para as menos frequentes). Nota-se que há alta sobreposição entre os conjuntos, naturalmente dominados por preposições, artigos e conjunções (consideradas *stopwords*). Não obstante, destaca-se a presença mais recorrente da palavra "como" nos textos gerados por máquina nos dois corpora. Em um primeiro olhar, essa listagem pode facilmente ser considerada irrelevante para a discriminação da origem da notícia, mas é importante lembrar que métodos clássicos de detecção de autoria consideram que essas palavras ajudam a identificar marcas estilísticas de autores, facilitando sua detecção (Mosteller e Wallace, 1964; Stamatatos, 2009), e talvez esse também seja o caso para a origem das notícias falsas.

Também foram levantadas as palavras mais frequentes após a remoção das *stopwords* (utilizando-se o NLTK), como pode ser visto na Tabela 3, buscando-se evidenciar mais claramente os conteúdos temáticos. Nesse caso, não foi possível observar discrepâncias entre as notícias.

Em uma análise mais refinada, fazendo-se um levantamento por classe de palavras, utilizando o etiquetador morfossintático porttagger (Silva et al., 2023), observam-se algumas informações lexicais interessantes:

- no Fake.br, há 49% de sobreposição de substantivos entre as notícias de humanos e máquinas e, no FakeTrueBR, esse valor cai para 35%, sendo que os lemas das palavras em comum que estão entre os mais frequentes nesses corpora são "presidente", "ano" e "governo";

- para verbos, esses valores são de 45% e 36%, respectivamente, sendo que os lemas das palavras em comum que estão entre os mais frequentes nesses corpora são "poder", "ter" , "afirmar", "dizer", "fazer", "revelar" e "receber";

[1] É importante ressaltar que, devido ao pré-processamento realizado, os valores aqui relatados podem divergir dos reportados por Silva et al. (2025).

Tabela 1: Características dos corpora.

Corpus	Núm. de notícias	Tokens/sentença	Tokens/notícia	Taxa type/token
Fake.br - Humano	3.600	15,58	191,14	0,5752
Fake.br - Máquina	3.600	20,57	368,78	0,5580
FakeTrueBR - Humano	1.791	17,19	163,17	0,5601
FakeTrueBR - Máquina	1.791	19,92	277,59	0,5846

Quadro 1: Exemplo de notícia falsa produzida por humano (à esquerda) e por máquina (à direita).

Humano	Máquina
Vergonha! Concessionária corta a energia do Maracanã por falta de pagamento. O consórcio Maracanã/Odebrecht deu um "calote" na Light. Desde setembro do ano de 2016, as contas de energia não são pagas. A dívida total ultrapassa R$ 3 milhões de reais. Na última copa do mundo, que foi realizada no Brasil, foram gastos aproximadamente R$ 1,3 bilhão no estádio. Nossos políticos, além de roubarem descaradamente o dinheiro público, agora querem tirar os momentos de lazer dos brasileiros [...] no caso o futebol!	Urgente: Luz do Maracanã cortada devido a suposta dívida milionária! A concessionária responsável pelo fornecimento de energia, Light Rio, anunciou que o corte foi necessário após meses de inadimplência. Segundo fontes exclusivas, a dívida teria ultrapassado R$ 5 milhões, envolvendo não apenas o estádio, mas também áreas adjacentes que não pagam suas contas desde a Olimpíada de 2016. A situação teria se agravado com a nova administração, que assumiu o controle em novembro e não teria honrado os compromissos financeiros. A prefeitura do Rio de Janeiro, questionada sobre o caso, não se pronunciou até o momento, gerando especulações sobre a gestão e o futuro de eventos no icônico estádio.

Tabela 2: Dez palavras mais frequentes nos corpora.

Fake.br (Humano)	Fake.br (Máquina)	FakeTrueBR (Humano)	FakeTrueBR (Máquina)
a	a	a	a
com	as	as	as
de	com	com	com
do	**como**	da	**como**
e	de	de	de
em	e	e	e
foi	em	em	em
na	não	não	não
no	o	o	o
não	os	os	os

Tabela 3: Dez palavras mais frequentes nos corpora (sem *stopwords*).

Fake.br (Humano)	Fake.br (Máquina)	FakeTrueBR (Humano)	FakeTrueBR (Máquina)
acordo	afirmou	agora	agora
ainda	agora	anos	ainda
anos	ainda	bolsonaro	apenas
brasil	anos	brasil	bolsonaro
dilma	apenas	dia	brasil
disse	brasil	gente	enquanto
durante	corrupção	governo	especialistas
federal	decisão	lula	governo
governo	enquanto	não	lula
lula	federal	pessoas	mensagem

- para adjetivos, tem-se respectivamente 43% e 28%, sendo que os lemas das palavras em comum que estão entre os mais frequentes nesses corpora são "novo", "político" , "grande", "social", "brasileiro" e "exclusivo".

É importante esclarecer que esse tipo de levantamento lexical é certamente afetado pela época coberta pelas notícias dos corpora e pelos tópicos mais presentes na mídia nesses momentos, que não são exatamente iguais, fazendo com que essas informações mais específicas mostradas acima sejam mais interessantes como caracterização dos dados do que como traços discriminativos da origem das notícias. No Fake.br, que contém a informação temporal das notícias (ver Figura 1), verificou-se que aproximadamente 88% das notícias falsas foram publicadas nos anos de 2016 e 2017, enquanto que cerca de 90% das notícias verdadeiras foram publicadas em 2017 e 2018. Isso indica uma concentração em tópicos políticos relacionados ao período da Operação Lava Jato e das eleições presidenciais de 2018. Por outro lado, o estudo de Chavarro et al. (2023) informou que as notícias da FakeTrueBR foram publicadas entre 2017 e 2023. Portanto, o FakeTrueBR abrange um período temporal mais amplo, incluindo não apenas notícias políticas, mas também eventos como a pandemia da COVID-19, o que pode ter contribuído para uma maior diversidade temática. Nesse sentido, abordagens futuras de modelagem de tópicos podem ser interessantes de serem aplicadas nos corpora.

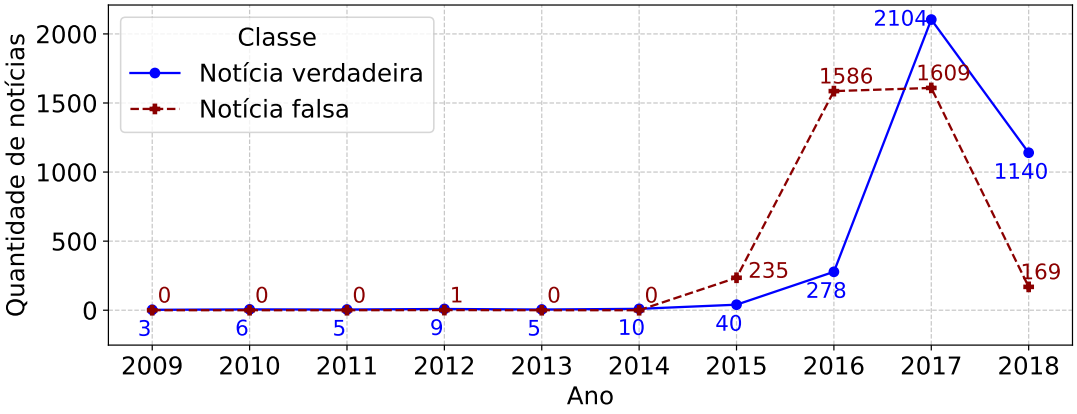


Figura 1: Distribuição temporal das notícias verdadeiras e falsas no corpus Fake.br.

Também foi aplicado o modelo SAGE (*Sparse Additive Generative Model*) proposto por Eisenstein et al. (2011) para identificar os termos mais distintivos dos textos humanos e de máquina. Diferentemente de abordagens baseadas em frequência absoluta, o SAGE modela a probabilidade de ocorrência de palavras como desvios esparsos em relação a uma distribuição de fundo comum, utilizando regularização L1 para induzir esparsidade nos parâmetros. Essa característica torna o modelo robusto contra ruído estatístico, permitindo que apenas os termos verdadeiramente distintivos apresentem valores significativos. Antes do cálculo, os textos passaram por normalização: conversão para minúsculas e remoção de acentos e cedilha. A distintividade de cada termo é calculada como a diferença entre os desvios específicos de cada classe, de modo que valores positivos indicam termos mais característicos de textos humanos, enquanto valores negativos indicam termos mais frequentes em textos gerados por LLM. A magnitude do valor reflete o grau de distintividade do termo. Na análise feita neste estudo, foram considerados tanto unigramas quanto bigramas para capturar padrões lexicais e colocacionais.

A Tabela 4 apresenta os dez termos mais distintivos para cada grupo. Os resultados revelam padrões morfossintáticos acerca dos textos gerados por máquinas, como a predominância de advérbios e adjetivos intensificadores. Esse contraste sugere que o LLM tende a produzir textos com maior carga de modificadores.

3.2 Tamanho de Palavras e de Sentenças

Investigou-se também a complexidade lexical dos textos por meio da contagem de sílabas. A escolha dessa métrica é motivada por trabalhos anteriores

Tabela 4: Termos mais distintivos segundo o SAGE (valores positivos indicam termos típicos de textos humanos; valores negativos, de textos gerados por LLM).

Humano		**Máquina**	
Termo	SAGE	Termo	SAGE
voce quiser	+7,36	fontes anonimas	−6,54
com gente	+7,32	reviravolta surpreendente	−5,88
org entre	+7,32	anonimas	−5,74
quiser sugerir	+7,31	fontes internas	−5,71
gente pelo	+7,28	fontes nao	−5,56
span	+7,14	revelacao bombastica	−5,43
site facebook	+7,13	enfatizando	−5,31
ao boatos	+7,07	acalorados	−5,29
esse artigo	+6,42	debates acalorados	−5,27
artigo uma	+6,20	cruciais	−5,23

que demonstraram sua relevância na distinção entre textos humanos e gerados por máquina. O trabalho de Shah et al. (2023) mostrou que características como contagem de sílabas e comprimento de palavras permitem distinguir esses textos com precisão de 93%. Além disso, Horne e Adalı (2017) identificaram que notícias falsas humanas tendem a utilizar vocabulário mais simples, enquanto Muñoz-Ortiz et al. (2024) observaram que textos gerados por LLMs apresentam distribuições de comprimento de sentença menos dispersas do que textos humanos. Essas evidências sugerem que métricas de complexidade silábica podem capturar diferenças estilísticas discriminativas entre os dois grupos. A Tabela 5 apresenta os resultados agregados dos dois corpora. As diferenças entre os grupos foram avaliadas por meio do teste t de Student, reportando-se o valor da estatística (t), o p-valor e o tamanho de efeito medido pelo d de Cohen.

Nota-se que a máquina produz, em média, palavras significativamente mais longas do que os humanos (2,39 vs 2,17 sílabas por palavra). Essa

Tabela 5: Complexidade silábica nos corpora.

Métrica	Humano	Máquina	*t*	*p*	*d* de Cohen
Sílabas/sentença	44,91 ± 32,97	52,48 ± 9,93	−16,14	< 0,001	0,31 (pequeno)
Sílabas/palavra	2,17 ± 0,14	2,39 ± 0,09	−97,06	< 0,001	1,87 (grande)

diferença apresenta tamanho de efeito grande (d = 1,87), indicando que os textos gerados por LLM tendem a utilizar palavras com maior complexidade silábica, o que sugere um vocabulário mais elaborado.

Observa-se que os textos humanos apresentam maior variabilidade na extensão das sentenças (desvio padrão de 32,97 sílabas por sentença, contra 9,93 para a máquina). Esse resultado se alinha com Muñoz-Ortiz et al. (2024), que, ao compararem textos jornalísticos em inglês produzidos por humanos e por LLMs, observaram que textos humanos apresentam distribuições de comprimento de sentença mais dispersas e maior variedade vocabular, sugerindo uma maior padronização estilística nos textos gerados automaticamente.

3.3 Classes Gramaticais

Com base na etiquetagem do porttagger (Silva et al., 2023), analisou-se comparativamente a distribuição de classes gramaticais entre os textos de humanos e de máquina. A motivação para essa análise reside nos resultados obtidos pelo estudo de Zhou et al. (2023), no qual identificou-se que textos noticiosos gerados por IA tendem a ser mais formais e analíticos, com menor uso de expressões informais e maior emprego de palavras emocionais. Essas evidências sugerem que a distribuição de classes gramaticais pode revelar padrões estilísticos distintos entre os dois grupos.

A Figura 2 mostra a proporção de cada classe gramatical nos corpora agregados. A proporção é calculada pelo número de ocorrências de cada classe dividido pelo número total de ocorrências de todas as classes em cada conjunto. É importante avaliar a proporção (e não a frequência absoluta), pois os textos têm tamanhos variados.

Observa-se, por exemplo, que os humanos usam proporcionalmente mais verbos (VERB e AUX), pronomes (PRON) e numerais (NUM), enquanto as máquinas fazem maior uso de substantivos (NOUN), nomes próprios (PROPN) e adjetivos (ADJ). O maior uso de adjetivos pela máquina (que parece ser a diferença mais relevante) está alinhado com os achados de Zhou et al. (2023), segundo os quais notícias falsas em inglês produzidas por máquina têm maior teor emocional (o que pode ser sinalizado pelos adjetivos).

Para verificar a significância estatística das diferenças, aplicou-se o teste qui-quadrado de independência. Embora o teste indique diferenças significativas na distribuição das classes (χ^2 = 21.579,28; $p < 0{,}001$), o V de Cramer (0,082) revela que a magnitude dessa associação é negligenciável. Isso sugere que, apesar das diferenças, a estrutura morfossintática dos textos produzidos por humanos e por máquinas é bastante similar.

3.4 Sintaxe

Para uma análise das diferenças sintáticas entre os textos, extraíram-se regras de formação de sentenças a partir das anotações de dependência no formato *Universal Dependencies* (de Marneffe et al., 2021), obtidas pelo uso do portparser (Lopes e Pardo, 2024), um analisador sintático automático para o português brasileiro. A análise sintática é motivada pelo estudo de Muñoz-Ortiz et al. (2024), cujas observações mostraram que textos humanos apresentam maior variedade no uso de tipos de dependência e constituintes sintáticos em comparação com textos gerados por LLMs. Adicionalmente, o trabalho de Shah et al. (2023) demonstrou que a estrutura de sentenças é um atributo relevante para a distinção entre textos humanos e de máquina. Assim, a extração de regras sintáticas permite investigar se essa menor diversidade estrutural também se manifesta no contexto específico de notícias falsas em português.

Cada regra representa um padrão estrutural que combina a classe gramatical de um *token* com as relações de dependência existentes com seus dependentes, em um formato que estende o formalismo discutido em Junior e Vale (2025). Por exemplo, a regra `NOUN(DET/det, *)` indica uma estrutura em que um substantivo (`NOUN`) tem um dependente que o precede na sentença e que é um determinante (`DET`), associado a ele pela relação de dependência *det*, enquanto a regra `*(VERB)` indica que o verbo é a raiz (*root*) da sentença.

As Tabelas 6 e 7 apresentam as estatísticas das

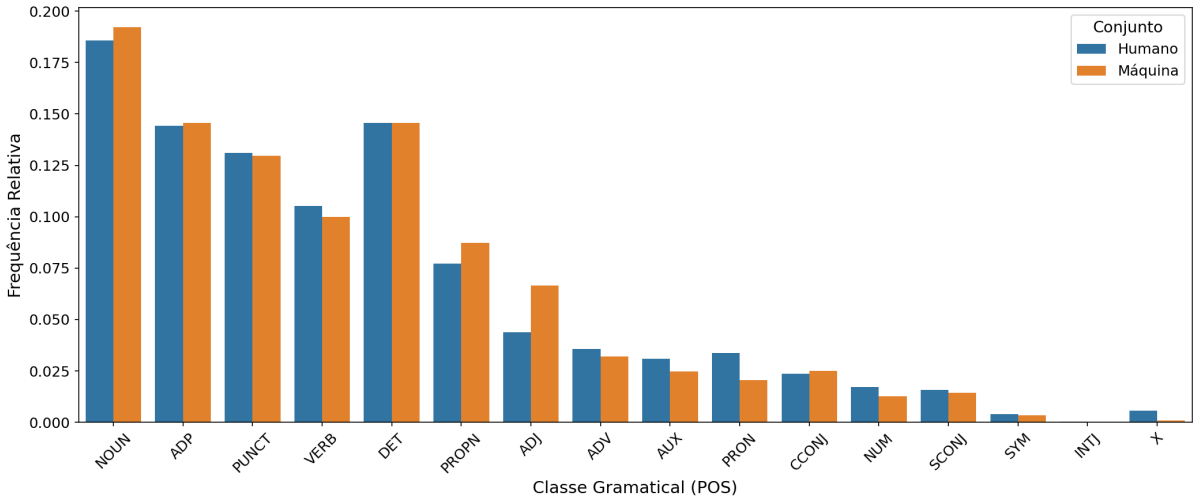


Figura 2: Comparação por classe gramatical nas notícias falsas produzidas por humanos e por máquinas.

regras extraídas. A coluna "Total de regras" indica a quantidade total de regras extraídas, contabilizando as repetições de uma mesma regra dentro de cada sentença. A coluna "Regras únicas" contabiliza apenas as regras distintas, sem repetição. De forma análoga, "Sentenças/texto" é a média de sentenças por notícia e "Regras/texto", a média de regras por notícia, enquanto "Únicas/texto" indica a média de regras distintas por notícia.

Tabela 6: Regras gramaticais – números absolutos.

Origem	Sentenças	Total de regras	Regras únicas
Humano	50.444	1.595.655	60.177
Máquina	79.170	2.938.295	57.658

Tabela 7: Regras gramaticais – médias.

Origem	Sentenças/texto	Regras/texto	Únicas/texto
Humano	9,36	296,04	74,11
Máquina	14,69	545,04	116,36

Observa-se que, embora a máquina produza mais regras por sentença (37,11 contra 31,63) e mais regras por notícia falsa (545,04 contra 296,04), o número total de regras únicas no corpus é ligeiramente menor (57.658 contra 60.177). Isso sugere maior repetição de padrões sintáticos nos textos gerados automaticamente, indicando uma menor diversidade sintática. Esse achado se alinha com Muñoz-Ortiz et al. (2024), que observou que humanos apresentam maior variedade no uso de tipos de dependência e constituintes sintáticos.

Para verificar a significância estatística dessas diferenças, aplicou-se o teste t de Student, cujos resultados estão na Tabela 8. Ambas as diferenças são estatisticamente significativas ($p < 0{,}001$), porém com tamanho de efeito pequeno (d = 0,27 e d = 0,38, respectivamente). Novamente, nota-se que os textos humanos apresentam maior variabilidade (desvio padrão de 24,62 contra 15,80 para regras por sentença), reforçando o padrão de maior homogeneidade estilística nos textos de máquina.

Para identificar regras mais discriminativas, aplicou-se uma adaptação da técnica TF-IDF (*Term Frequency-Inverse Document Frequency*) sobre as regras extraídas, tratando cada texto como um documento e cada regra como um termo. Foram consideradas apenas regras que possuíssem pelo menos um dependente, excluindo-se regras simples que só introduzissem itens léxicos. Em seguida, comparou-se a distribuição dos *scores* TF-IDF entre os grupos humano e máquina por meio do teste de Mann-Whitney, calculando-se também o d de Cohen para avaliar o tamanho do efeito. As Tabelas 9 e 10 apresentam as regras mais discriminativas para cada origem, acompanhadas de exemplos de ocorrência. A coluna "Diff" representa a diferença na taxa média de uso da regra entre os grupos (valores positivos indicam regras mais frequentes em textos humanos, enquanto valores negativos indicam regras mais frequentes em textos de máquina).

Observa-se que as regras discriminantes de humanos apresentam tamanhos de efeito predominantemente pequenos (d entre 0,22 e 0,56), enquanto as regras discriminantes de textos de máquina apresentam efeitos de médios a grandes (d entre 0,58 e 1,30). Essa assimetria corrobora os achados anteriores sobre a maior homogeneidade estilística dos

Tabela 8: Testes estatísticos para regras de formação de sentenças.

Métrica	**Humano**	**Máquina**	t	p	d **de Cohen**
Regras/sent.	$31{,}63 \pm 24{,}62$	$37{,}11 \pm 15{,}80$	$-48{,}83$	$< 0{,}001$	0,27 (pequeno)
Regras únicas/sent.	$16{,}85 \pm 7{,}80$	$19{,}41 \pm 5{,}51$	$-69{,}21$	$< 0{,}001$	0,38 (pequeno)

Tabela 9: Regras discriminantes de textos humanos.

Regra	**Diff**	p	d	**Exemplo**
PUNCT(PUNCT/punct, *, PUNCT/punct, PUNCT/punct)	+0,012	$< 0{,}001$	0,56	*No seu governo fez-se o diabo **[...]** no seu partido*
ADV(*, NUM/appos)	+0,006	$< 0{,}001$	0,30	*O Vaticano iniciou **ontem** (**18**) o julgamento de dois ex-dirigentes do hospital infantil da Santa Sé, em Roma.*
SYM(ADP/case, DET/det, *)	+0,005	$< 0{,}001$	0,29	*Além **dos R$** 3,2 milhões para o PT, Bené disse que pegou R$ 250 mil em dinheiro vivo.*
VERB(SCONJ/mark, PRON/nsubj, *, VERB/xcomp, PUNCT/punct)	+0,005	$< 0{,}001$	0,29	***Se alguém tiver** que **cair,** esse alguém vai ser o Temer.*
ADV(ADP/case, *, NUM/appos)	+0,005	$< 0{,}001$	0,25	*Na manhã **de hoje** (**17**) o neto do ex-presidente também cometeu suicídio aos 61 anos.*
VERB(*, PUNCT/punct)	+0,005	$< 0{,}001$	0,25	***Esqueça!***
PRON(*, VERB/acl:relcl)	+0,005	$< 0{,}001$	0,22	*Enfim, veja aos 32:40 Dr. Rey espinafrando Bolsonaro: **Quem dá** brechas não pode reclamar que os outros aproveitem, não é mesmo?*
VERB(*, NOUN/obj, SYM/obl)	+0,005	$< 0{,}001$	0,27	*Ele estava com a mulher e queria **vender** os 2 **ingressos** por **R$** 2 mil.*
NOUN(ADP/case, DET/det, *, NOUN/conj, NOUN/conj)	+0,004	$< 0{,}001$	0,26	*O garoto foi encaminhado para o hospital com ferimentos **nos braços**, **pernas** e **rosto**.*
NOUN(ADP/case, DET/det, *)	+0,004	$< 0{,}001$	0,50	*Essas pessoas que me expulsaram não servem **ao país**.*

Tabela 10: Regras discriminantes de textos gerados por máquina.

Regra	**Diff**	p	d	**Exemplo**
VERB(PUNCT/punct, *, NOUN/obj)	$-0{,}021$	$< 0{,}001$	1,30	*PHA (como é conhecido nas redes sociais) tem uma obsessão doentia**: Prender** os **donos** da Rede Globo.*
VERB(PUNCT/punct, *, VERB/ccomp)	$-0{,}016$	$< 0{,}001$	0,79	*Esse processo nos levou a explorar a respiração e sua presença dentro de nossas vidas**, mostrando** que **podemos** estar sem comida enquanto houver ar.*
ADV(ADV/advmod, *)	$-0{,}014$	$< 0{,}001$	0,71	*Fidel **não só** representa uma Cuba diferente, mas agora parece que ele transcendeu a própria morte!*
NOUN(DET/det, *, ADJ/amod, NOUN/nmod)	$-0{,}014$	$< 0{,}001$	0,68	*No Brasil, uma bomba-relógio judicial pode definir **o futuro político** do **país**.*
PROPN(PUNCT/punct, *, PUNCT/punct)	$-0{,}013$	$< 0{,}001$	0,66	*A decisão também beneficia Frederico Pacheco, primo do senador, e Mendherson Souza Lima, ex-assessor parlamentar de Zeze Perrella **(PMDB-MG)**.*
NOUN(*, ADJ/amod)	$-0{,}013$	$< 0{,}001$	0,73	*Definindo Bolsonaro como um "homem autêntico e corajoso", Bivar promete **surpresas positivas**.*
ADV(*, PRON/fixed, PUNCT/punct)	$-0{,}013$	$< 0{,}001$	0,67	***Enquanto isso,** a falta de documentação adequada impede famílias como a de Reginaldo de acessar benefícios sociais, como o Bolsa Família.*
NOUN(*, ADJ/amod, NOUN/nmod)	$-0{,}013$	$< 0{,}001$	0,62	*Uma megaconferência nos EUA está causando alvoroço ao reunir **figuras controversas** da **política** brasileira.*
VERB(NOUN/nsubj, *, VERB/ccomp, PUNCT/punct)	$-0{,}012$	$< 0{,}001$	0,58	***Rumores indicam** que ele também **enfrenta** dissidência dentro do chavismo, com Rafael Ramírez emergindo como uma possível alternativa.*
NOUN(ADP/case, *, ADJ/amod)	$-0{,}012$	$< 0{,}001$	0,77	*Eles alegam que Bolsonaro é o oposto dos ideais **de liberdade econômica** e comportamental que desejavam promover.*

textos gerados automaticamente.

Entre as regras discriminantes de humanos, destacam-se estruturas de pontuação complexas, como `PUNCT(PUNCT/punct, *, PUNCT/punct, PUNCT/punct)`, que refletem o uso de reticências ou múltiplos sinais de pontuação típicos de um estilo mais informal. Também se observa maior uso de estruturas com símbolos monetários, como `SYM(ADP/case, DET/det, *)`, e sentenças curtas imperativas, como `VERB(*, PUNCT/punct)`.

As regras discriminantes da máquina revelam maior uso de modificadores adjetivais, como `NOUN(*, ADJ/amod)` e `NOUN(DET/det, *, ADJ/amod, NOUN/nmod)`, corroborando a análise de classes gramaticais. A regra com maior poder discriminativo, `VERB(PUNCT/punct, *, NOUN/obj)`, apresenta efeito grande (d = 1,30), indicando uma forte tendência da máquina em produzir estruturas com objeto direto. A presença de complementos oracionais, evidenciada por `VERB(PUNCT/punct, *, VERB/ccomp)` e `VERB(NOUN/nsubj, *, VERB/ccomp, PUNCT/punct)`, sugere construções sintáticas sofisticadas nos textos de máquina.

4 Discussão e Considerações Finais

As análises realizadas revelam diferenças consistentes entre as notícias falsas produzidas por humanos

e por máquinas, embora com magnitudes variadas. Um padrão recorrente em todas as análises foi a maior variabilidade dos textos humanos. Por exemplo, os desvios padrão para sílabas por sentença (32,97 contra 9,93), regras sintáticas por sentença (24,62 contra 15,80) e regras únicas por sentença (7,80 contra 5,51) foram consistentemente maiores nos textos humanos. Esses resultados reforçam a maior homogeneidade estilística dos textos gerados por máquina. A distribuição de classes gramaticais foi o único caso em que se observou bastante similaridade entre textos de humanos e de máquinas.

Algumas ressalvas e observações são relevantes de serem feitas com relação aos resultados obtidos. Inicialmente, é importante considerar que o método de geração das notícias falsas pela máquina certamente tem impacto nos resultados. Nos corpora utilizados, a máquina utilizou as notícias verdadeiras como base para a geração das falsas. Outros caminhos possíveis incluem a geração sem o uso das notícias verdadeiras como base ou o uso das notícias falsas escritas por humanos como referência. O grande modelo de língua utilizado, o Sabiá-3, também pode influenciar os resultados, sendo sua escolha motivada pelo melhor alinhamento com a língua portuguesa e sua cultura associada. Do ponto de vista de processamento automático, o etiquetador morfossintático e o *parser* de dependências utilizados são do estado da arte para o português brasileiro, mas, como é usual na área, não são perfeitos. Erros pontuais de anotação são esperados, mas não devem alterar os resultados observados. Também é importante destacar que esses sistemas seguem o modelo *Universal Dependencies*, tornando esta investigação mais relevante e, se houver interesse, mais facilmente replicável e comparável com outras línguas.

Outra característica interessante deste estudo refere-se a uma hipótese subjacente, de que discriminar notícias falsas de humanos e de máquinas é diferente da tarefa mais geral de discriminar textos escritos por humanos (não importando se contêm informações verídicas ou não) de textos produzidos por máquina. Conforme discutido anteriormente neste artigo, a literatura aponta nessa direção. Alguns estudos, inclusive, ressaltam que estratégias de escrita diferentes (muitas vezes inconscientes) são utilizadas em notícias falsas, o que sinaliza a relevância de estudos metodológicos diferenciados para esses casos. Além dessa questão, neste artigo não se diferenciaram os casos relacionados de *misinformation*, *disinformation* and *malinformation*, como definidos por Wardle e Derakhshan (2017), uma vez que os corpora existentes não fazem essa distinção. Porém, esses conceitos consistem em outro ponto relevante que pode interferir na hipótese assumida e influenciar nos resultados.

Por fim, trabalhos futuros incluem a possível continuidade da categorização das notícias, abordando-se os níveis mais desafiadores da semântica e do discurso. Também se pretende avaliar métodos de classificação automática da origem das notícias. Espera-se que a análise apresentada neste artigo possa subsidiar esses métodos e decisões de projeto relacionadas.

Agradecimentos

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES). Ele também contou com o apoio da Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP; processo #2024/17834-6) e do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq; processo #444933/2024-7).

Referências

Hugo Abonizio, Thales Sales Almeida, Thiago Laitz, Roseval Malaquias Junior, Giovana Kerche Bonás, Rodrigo Nogueira, e Ramon Pires. 2024. Sabiá-3 technical report. *Preprint*, arXiv:2410.12049.

Ankit Aich, Souvik Bhattacharya, e Natalie Parde. 2022. Demystifying neural fake news via linguistic feature-based interpretation. Em *Proceedings of the 29th International Conference on Computational Linguistics*, páginas 6586–6599, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Ramy Baly, Georgi Karadzhov, Jisun An, Haewoon Kwak, Yoan Dinkov, Ahmed Ali, James Glass, e Preslav Nakov. 2020. What was written vs. who read it: News media profiling using text analysis and social media context. Em *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, páginas 3364–3374. Association for Computational Linguistics.

Fabrizio Barnabò, Federico Siciliano, Carlos Castillo, Stefano Leonardi, e Preslav Nakov. 2022. Deep learning for cross-lingual news stance detection. Em *Proceedings of the ACM Web Conference 2022*, páginas 1094–1103. ACM.

Tony Berber Sardinha. 2024. Ai-generated vs human-authored texts: A multidimensional comparison. *Applied Corpus Linguistics*, 4(1):100083.

Juan Pablo Chavarro, Jonata Tyska Carvalho, Tarlis Tortelli Portela, e Jonathan Cardoso Silva. 2023. FakeTrueBR: Um corpus brasileiro de notícias falsas. Em *Anais da XVIII Escola Regional de Banco de Dados*, páginas 108–117, Porto Alegre. SBC.

Canyu Chen e Kai Shu. 2024. Combating misinformation in the age of LLMs: Opportunities and challenges. *AI Magazine*, 45(3):354–368.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, e Daniel Zeman. 2021. Universal dependencies. *Computational Linguistics*, 47(2):255–308.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, e Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. Em *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, páginas 4171–4186. Association for Computational Linguistics.

Jacob Eisenstein, Amr Ahmed, e Eric P. Xing. 2011. Sparse additive generative models of text. Em *Proceedings of the 28th International Conference on Machine Learning*, ICML'11, páginas 1041–1048, Madison, WI, USA. Omnipress.

Jorge Fernández García e Isabel Segura-Bedmar. 2024. Human after all: Using transformer based models to identify automatically generated text. Em *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2024)*, volume 3756 de *CEUR Workshop Proceedings*. CEUR-WS.org.

Lucas Graves e Federica Cherubini. 2016. The rise of fact-checking sites in Europe. Relatório técnico, Reuters Institute for the Study of Journalism, Oxford, UK.

Benjamin D. Horne e Sibel Adalı. 2017. This just in: Fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news. Em *Proceedings of the International AAAI Conference on Web and Social Media*, volume 11, páginas 759–766.

Zhiwei Jin, Juan Cao, Han Guo, Yongdong Zhang, e Jiebo Luo. 2017. Multimodal fusion with recurrent neural networks for rumor detection on microblogs. Em *Proceedings of the 25th ACM International Conference on Multimedia*, páginas 795–816. ACM.

Isaac Souza de Miranda Junior e Oto Araújo Vale. 2025. Dependência: o conceito e as gramáticas. *Linguamática*, 17(2):71–84.

Lucelene Lopes e Thiago Alexandre Salgueiro Pardo. 2024. Towards portparser – a highly accurate parsing system for Brazilian Portuguese following the Universal Dependencies framework. Em *Proceedings of the 16th International Conference on Computational Processing of Portuguese (PROPOR 2024)*, páginas 399–409, Santiago de Compostela, Galicia/Spain. Association for Computational Linguistics.

Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D. Manning, e Chelsea Finn. 2023. DetectGPT: Zero-shot machine-generated text detection using probability curvature. Em *Proceedings of the 40th International Conference on Machine Learning*, volume 202 de *Proceedings of Machine Learning Research*, páginas 24950–24962. PMLR.

Rafael A. Monteiro, Roney L. S. Santos, Thiago A. S. Pardo, Tiago A. de Almeida, Evandro E. S. Ruiz, e Oto A. Vale. 2018. Contributions to the study of fake news in Portuguese: New corpus and automatic detection results. Em *Computational Processing of the Portuguese Language*, volume 11122 de *Lecture Notes in Computer Science*, páginas 324–334, Cham. Springer.

Frederick Mosteller e David L. Wallace. 1964. *Inference and Disputed Authorship: The Federalist*. Addison-Wesley Series in Behavioral Science: Quantitative Methods. Addison-Wesley, Reading, MA.

Alberto Muñoz-Ortiz, Carlos Gómez-Rodríguez, e David Vilares. 2024. Contrasting linguistic patterns in human and LLM-generated news text. *Artificial Intelligence Review*, 57(10):265.

Verónica Pérez-Rosas, Bennett Kleinberg, Alexandra Lefevre, e Rada Mihalcea. 2018. Automatic detection of fake news. Em *Proceedings of the 27th International Conference on Computational Linguistics*, páginas 3391–3401, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Ramon Pires, Hugo Abonizio, Thales Sales Almeida, e Rodrigo Nogueira. 2023. Sabiá: Portuguese large language models. Em *Intelligent Systems*, volume 14197 de *Lecture Notes in Computer Science*, páginas 226–240, Cham. Springer.

Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, e Yejin Choi. 2017. Truth of varying shades: Analyzing language in fake news and political fact-checking. Em *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, páginas 2931–2937, Copenhagen, Denmark. Association for Computational Linguistics.

Aditya Shah, Mohand Boughanem, e Gabriella Pasi. 2023. Feature-based detection of machine-generated text. *Expert Systems with Applications*, 228:120321.

Emanuel Huber Silva, Thiago Alexandre Salgueiro Pardo, e Norton Trevisan Roman. 2023. Etiquetagem morfossintática multigênero para o português do Brasil segundo o modelo "Universal Dependencies". Em *Anais do XIV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana (STIL)*, páginas 63–73, Belo Horizonte, Brasil. Sociedade Brasileira de Computação.

Renato M. Silva, Roney L. S. Santos, Tiago A. Almeida, e Thiago A. S. Pardo. 2020. Towards automatically filtering fake news in Portuguese. *Expert Systems with Applications*, 146:113199.

Renato Moraes Silva, Hazem Amamou, Lucca Baptista Silva Ferraz, Fabio Kauê Araujo da Silva, e Anderson Raymundo Avila. 2025. Fake news detection in Portuguese under large language model-generated content. *Journal of the Brazilian Computer Society*, 31(1):1149–1166.

Renato Moraes Silva, Roney Lira de Sales Santos, e Thiago Alexandre Salgueiro Pardo. 2024. Detecção automática de notícias falsas. Em Helena de Medeiros Caseli e Maria das Graças Volpe Nunes, editores, *Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português*, 3 edição, capítulo 27. BPLN.

Efstathios Stamatatos. 2009. A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology*, 60(3):538–556.

Jinyan Su, Claire Cardie, e Preslav Nakov. 2024a. Adapting fake news detection to the era of large language models. Em *Findings of the Association for Computational Linguistics: NAACL 2024*, páginas 1473–1490, Mexico City, Mexico. Association for Computational Linguistics.

Jinyan Su, Claire Cardie, e Preslav Nakov. 2024b. Adapting fake news detection to the era of large language models. Em *Findings of the Association for Computational Linguistics: NAACL 2024*, páginas 1473–1490, Mexico City, Mexico. Association for Computational Linguistics.

Jinyan Su, Terry Yue Zhuo, Javid Mansurov, Di Wang, e Preslav Nakov. 2023. Fake news detectors are biased against texts generated by large language models. Em *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, páginas 10276–10288. Association for Computational Linguistics.

Claire Wardle e Hossein Derakhshan. 2017. *INFORMATION DISORDER: Toward an interdisciplinary framework for research and policy making*.

Jiawei Zhou, Yixuan Zhang, Qianni Luo, Andrea G. Parker, e Munmun De Choudhury. 2023. Synthetic lies: Understanding AI-generated misinformation and evaluating algorithmic and human solutions. Em *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, páginas 1–20, New York, NY, USA. ACM.

Xinyi Zhou e Reza Zafarani. 2020. A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Computing Surveys*, 53(5):1–40.

Exploração de métodos simbólicos para detecção de emoções para o português

Stephanie Briere Americo
Núcleo Interinstitucional de Linguística Computacional (NILC)
Instituto de Ciências Matemáticas e de Computação (ICMC)
Universidade de São Paulo
stephanieb.americo@usp.br

Thiago Alexandre Salgueiro Pardo
Núcleo Interinstitucional de Linguística Computacional (NILC)
Instituto de Ciências Matemáticas e de Computação (ICMC)
Universidade de São Paulo
taspardo@icmc.usp.br

Resumo

Este trabalho investiga métodos simbólicos para a detecção de emoções em textos em português, considerando múltiplos córpus, domínios e diferentes configurações de pré-processamento. Os resultados mostram grande variação no desempenho absoluto entre domínios, mas estabilidade no desempenho relativo entre os métodos, evidenciando a influência das propriedades do córpus e o gradiente entre complexidade e interpretabilidade. A inclusão da classe neutra tende a degradar o desempenho ao aumentar a ambiguidade e, frequentemente, o desbalanceamento entre classes, enquanto um pré-processamento mais extensivo beneficia especialmente abordagens simbólicas. A análise qualitativa indica que parte dos erros decorre de ambiguidades linguísticas, do grande espaço para subjetividade no processo de anotação e das próprias nuances emocionais, reforçando a importância de avaliações comparativas multi-domínio.

1 Introdução

A detecção automática de emoções em textos é uma tarefa relevante do Processamento de Linguagem Natural, com aplicações em mídias sociais, sistemas de recomendação, educação, atendimento ao cliente e monitoramento de saúde mental (Mohammad, 2016). Diferentemente da análise de sentimentos baseada apenas em polaridade, essa tarefa busca identificar estados afetivos mais complexos, como alegria, tristeza, raiva e medo.

A ambiguidade da linguagem natural e a subjetividade das expressões emocionais dificultam tanto a interpretação humana quanto a análise automática de emoções. As amostras da Tabela 1 ilustram esse fenômeno: na amostra 1, um *review* de filme combina uma opinião positiva com uma ressalva negativa, evidenciando a subjetividade da avaliação; na amostra 2, um *review* de jogo de horror utiliza emoções tipicamente negativas, como o medo, para expressar uma experiência positiva. Esses exemplos também mostram que a polaridade textual não está necessariamente alinhada às emoções expressas, pois ambas são fortemente dependentes do contexto e, embora relacionadas, não são mutuamente determinantes.

	Texto
1	O filme é muito bom, mas não recomendaria para minha família.
2	Assustador e viciante, não consegui parar de jogar.

Tabela 1: Exemplos de textos nos quais a detecção de emoções não é trivial.

Modelos neurais de larga escala têm alcançado desempenho superior ao de métodos clássicos de aprendizado de máquina (Devlin et al., 2019; Maruf et al., 2024). No entanto, por serem treinados sobre grandes volumes de dados potencialmente enviesados, esses modelos podem herdar preconceitos humanos e produzir decisões injustas. Além disso, sua natureza opaca dificulta a interpretação, auditoria e responsabilização dos sistemas automáticos. Em contraste, métodos simbólicos oferecem transparência, controle explícito do conhecimento e previsibilidade do comportamento do modelo, características historicamente centrais para a Inteligência Artificial e ainda cruciais em aplicações que exigem resultados interpretáveis por humanos.

É neste contexto que abordagens simbólicas permanecem relevantes, especialmente quando interpretabilidade, baixo custo computacional e facilidade de adaptação a novos domínios são requisitos centrais. Em áreas sensíveis, como saúde mental, educação e sistemas de tomada de decisão automatizada, compreender por que uma emoção foi atribuída é uma exigência ética tão importante quanto o desempenho preditivo (Guidotti et al., 2018; Ribeiro et al., 2016).

Este estudo integra um esforço mais amplo que busca avaliar a detecção de emoções em português de forma abrangente, considerando todos os córpus

Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)
Vol. 2, pages 159–169

publicamente disponíveis e diferentes esquemas de anotação, sem restringir o conjunto de emoções ou o paradigma dos modelos de classificação avaliados. O presente artigo constitui um recorte inicial dessa investigação, com foco em métodos simbólicos.

Consideramos desafios centrais da tarefa, como subjetividade e ambiguidade linguística, cenários *multilabel*, desbalanceamento de classes, variação entre esquemas de anotação, adaptação a domínios e escassez de recursos (Maruf et al., 2024). As contribuições são: (i) uma comparação empírica entre métodos com diferentes níveis de interpretabilidade em múltiplos córpus; (ii) uma análise qualitativa da interpretabilidade, com foco nas regras e estruturas aprendidas; e (iii) uma discussão de limitações e oportunidades de melhoria.

2 Trabalhos relacionados

Em nossa revisão da literatura, não identificamos trabalhos em português que utilizem abordagens simbólicas para a detecção de emoções. Esta seção apresenta estudos relevantes que exploram outras abordagens para o português, incluindo trabalhos baseados em córpus também utilizados neste estudo e descritos em mais detalhes na Seção 3, bem como a menção a outros trabalhos com abordagens semelhantes.

O estudo de da Silva (2020) avaliou classificadores SVM com *kernel* de árvore para detecção de emoções em textos em um cenário de transferência de domínio, utilizando um córpus com publicações de domínio livre para treinamento e um córpus reduzido com publicações sobre o mercado de ações (DANTEStocks) para teste. Foram treinados classificadores binários para pares de emoções opostas segundo a roda de Plutchik, alcançando Medida-F média ponderada de 0,56. Os resultados evidenciaram o impacto do tamanho e do desbalanceamento das classes, bem como a degradação do desempenho com a inclusão da classe neutra, indicando limitações de generalização entre domínios e dificuldades na classificação de textos ambivalentes.

Um modelo pré-treinado para o português (BERTimbau (Souza et al., 2020)) foi ajustado ao córpus GoEmotions-BR para detectar as seis emoções básicas de Ekman e a classe neutra (Oliveira and Sichman, 2024), seguindo procedimento análogo ao adotado no GoEmotions original em inglês. Embora os resultados não sejam diretamente comparáveis devido ao uso de córpus distintos, os valores de Medida-F apresentam ordem de magnitude e consistência semelhantes entre os modelos, com média geral de 0,57. Observou-se desempenho significativamente superior na detecção da classe "alegria" em relação a "nojo", atribuível principalmente à maior representatividade da primeira no treinamento e à sua expressão mais explícita em textos, enquanto "nojo" é menos frequente e mais dependente de nuances contextuais.

Outros trabalhos em português também investigam a detecção de emoções em textos, adotando abordagens semelhantes às descritas anteriormente. Hammes and de Freitas (2021) utilizou o modelo BERTimbau no córpus GoEmotions traduzido com o Google Tradutor, resultando em um recurso distinto do anteriormente mencionado (GoEmotions-BR). Diferentemente de trabalhos que mapeiam as emoções para o conjunto reduzido das seis emoções básicas de Ekman acrescidas da classe neutra, esse estudo considerou as 28 classes originais do GoEmotions. Os resultados obtidos foram compatíveis com os reportados para o modelo BERT original em inglês, alcançando Medida-F média de 0,48.

Dosciatti et al. (2013) aplicou classificadores Naive Bayes, kNN e SVM ao córpus Notícias Curtas, reportando que o SVM obteve o melhor desempenho, com acurácia de 0,61. Outros estudos relevantes incluem Cortiz et al. (2021), que também empregou modelos baseados em BERT, e Santos (2019), que comparou algumas das mesmas abordagens clássicas de aprendizado de máquina.

3 Conjuntos de dados

Os córpus utilizados foram selecionados após uma revisão abrangente dos recursos disponíveis para a língua portuguesa, obtidos por disponibilização pública ou por contato direto com seus autores. Alguns córpus descritos na literatura não foram considerados neste trabalho devido a restrições de acesso ou licenciamento, ou por redundância em relação a recursos já adotados. Estes córpus são descritos em Hammes and de Freitas (2021); Cortiz et al. (2021); Santos (2019); Duarte (2019). Essa escolha reflete limitações práticas de disponibilidade, preservando ainda diversidade de domínios, esquemas de anotação e modelos de emoção adotados.

O córpus **DANTEStocks**[1] contém 4.277 publi-

[1]Disponível em: https://www.kaggle.com/datasets/fernandojvdasilva/stock-tweets-ptbr-emotions.

cações em português da rede social *X*, coletadas em 2014 e relacionadas a ações do índice Ibovespa (da Silva et al., 2020). Os textos foram anotados de forma supervisionada com a classe neutra e as oito emoções primárias da roda de Plutchik (Plutchik, 2001), organizadas em quatro pares de opostos: confiança e nojo; alegria e tristeza; antecipação e surpresa; irritação e medo. O córpus apresenta forte desbalanceamento entre classes.

da Silva et al. (2020) disponibilizou também outro córpus com um subconjunto reduzido do primeiro, contendo apenas 334 textos que obtiveram concordância total entre anotadores, sendo majoritariamente composto por exemplos neutros. Para facilitar a apresentação dos resultados, denominamos esta variação do córpus **DANTEStocks Concordância**.

O córpus **Domínio Livre** é composto por 230.857 publicações da rede social *X*, coletadas sem restrição temática e anotadas automaticamente por supervisão distante, a partir de *hashtags* associadas às emoções de Plutchik. É um conjunto desbalanceado e não possui publicações "neutras" (da Silva, 2020).

Notícias Curtas[2] é um córpus composto por 1.750 notícias do jornal virtual *O Globo*, anotadas manualmente com a classe neutra e as seis emoções básicas de Ekman (Ekman and Keltner, 1997): alegria, tristeza, medo, surpresa, nojo e raiva. O conjunto é balanceado, com 250 textos por classe, e a anotação foi realizada por especialistas com resolução de divergências por consenso (Dosciatti et al., 2013).

O **GoEmotions-BR** (Oliveira and Sichman, 2024) deriva da tradução automática do córpus GoEmotions (Demszky et al., 2020), originalmente composto por um conjunto desbalanceado de 58 mil sentenças em inglês anotadas em 27 emoções ("alegria", "diversão", "entusiasmo", "luto", "tristeza", "raiva", "aborrecimento", entre outras) e neutro. A tradução foi realizada com o modelo *chatGPT3.5-turbo*.

Nos experimentos de Demszky et al. (2020), além da versão original do GoEmotions com 27 emoções granulares, foi proposta uma variante reorganizada em que as emoções são mapeadas para as 6 categorias básicas de Ekman, acrescidas da classe neutra (Oliveira and Sichman, 2024). Adotamos também essa variação (**GoEmotions-BR-Ekman**), pois ela permite comparar diretamente, em um mesmo domínio, o impacto de diferentes modelos de emoção sobre os resultados.

Com exceção do córpus Notícias Curtas, os córpus descritos nesta seção possuem anotação *multilabel*. Nesse cenário, cada amostra pode receber um ou mais rótulos de classes emocionais, como ilustrado na Tabela 2. Essa característica reflete de forma mais fiel a complexidade emocional humana, mas também aumenta a complexidade e a dificuldade da tarefa de detecção automática.

	Texto	***Rótulo***
1	VALE5 nao passa de 29,90	Antecipação
2	Algo me diz que vou ver a #PETR4 na casa dos 12 hoje ainda.. vamos aguardar.. :)	Confiança, Alegria, Antecipação

Tabela 2: Exemplo de amostra que recebeu apenas um rótulo de emoção e de amostra que foi anotada com múltiplos rótulos (*multilabel*). As amostras são do córpus DANTEStocks (da Silva, 2020).

É importante mencionar que não há consenso científico sobre a natureza e a delimitação das emoções humanas, e os modelos adotados na literatura baseiam-se em diferentes teorias e métodos (como expressões faciais em Ekman e *crowdsourcing* no GoEmotions). A classe "neutra" é ainda mais controversa, pois pressupõe a ausência de emoção, uma condição para a qual não há evidência empírica clara. Embora cada córpus apresente sua própria definição, essa classe tende a agrupar casos heterogêneos, como baixa intensidade emocional, ambiguidade ou ausência de pistas suficientes para a anotação. Além disso, frequentemente constitui a classe predominante em córpus desbalanceados, possivelmente por concentrar casos de difícil caracterização. Consequentemente, trata-se de uma categoria pouco definida, cuja utilização pode introduzir ruído e dificultar a interpretação e a modelagem das emoções.

A Tabela 3 resume os córpus utilizados neste trabalho, que diferem quanto ao domínio, esquema de anotação, modelo de emoções e balanceamento de classes.

4 Metodologia

Esta seção descreve os métodos avaliados e a configuração experimental adotada, fornecendo as informações necessárias para garantir a reprodutibilidade dos resultados. Detalhes adicionais, bem como esclarecimentos sobre decisões de implementação, podem ser obtidos mediante contato com os

[2]Disponível para solicitação em: https://www.ppgia.pucpr.br/~paraiso/mineracaodeemocoes/recursos.php.

córpus	*Modelo de emoções*	*Tamanho*	*Anotação*
DANTE	Plutchik	4.277	Superv.
DANTEConcordancia	Plutchik	334	Superv.
DominioLivre	Plutchik	230.857	Automat.
NoticiasCurtas	Ekman	1.750	Superv.
GoEmotionsBR	27 e neutro	∼58.000	Superv.
GoEmotionsBREkman	Ekman	∼58.000	Superv.

Tabela 3: Resumo dos córpus analisados

autores.

4.1 Métodos avaliados

Avaliamos cinco métodos de detecção de emoções em texto, em sua maioria classificadores simbólicos, com diferentes níveis de interpretabilidade. A seleção prioriza modelos consolidados, de fácil implementação e interpretação, permitindo uma comparação consistente entre abordagens simbólicas, supervisionadas e parcialmente interpretáveis.

O **ZeroR** é adotado como *baseline* por sempre prever a classe majoritária do conjunto de treinamento, estabelecendo um limite inferior de desempenho (Alnuaimi and Albaldawi, 2024).

Propomos um método simbólico baseado em regras e léxico (**LexicoR**), inspirado em Seal et al. (2020). O método utiliza o léxico emocional *Emocionário* (Ramos, 2021), no qual termos emocionais são manualmente identificados e anotados por especialistas com base em obras de referência da língua portuguesa. A partir desse recurso, o LexicoR atribui pontuações às emoções conforme a presença e a frequência de termos emocionais, produzindo decisões totalmente transparentes. O objetivo é avaliar o desempenho alcançável por uma abordagem puramente simbólica, baseada exclusivamente em conhecimento linguístico e regras manuais, e verificar se ela supera de forma consistente um *baseline* ingênuo (ZeroR) em múltiplos domínios.

Para cada *token* do texto que esteja presente no *Emocionário*, considera-se uma janela de contexto simétrica de quatro palavras anteriores e posteriores, na qual são identificados modificadores semânticos de negação, intensificação e redução. A pontuação base é 1; intensificadores multiplicam esse valor por 3, redutores o dividem por 3, e a negação inverte o sinal.

Diferentemente de Seal et al. (2020), não assumimos relações explícitas de oposição entre emoções. As pontuações são acumuladas por emoção ao longo do texto, e apenas as emoções dominantes — aquelas com pontuação máxima — são previstas. Em caso de empate, todas as emoções correspondentes são atribuídas, evitando a definição arbitrária de limiares para classificação *multilabel*. O Apêndice A detalha o funcionamento da solução.

A **árvore de decisão** é um modelo supervisionado que aprende regras hierárquicas por meio de testes sobre atributos (Murthy, 1998). Trata-se de um método simbólico, pois pode ser expresso como um conjunto de regras, sendo sua interpretabilidade maximizada na representação em forma de árvore: cada caminho da raiz até uma folha corresponde a uma regra explícita. Como o modelo não oferece suporte nativo à classificação *multilabel*, adotamos a estratégia um-contra-todos, induzindo uma árvore independente para cada emoção.

O **RIPPER** induz regras proposicionais diretamente de dados rotulados, expressas como condições lógicas simples e facilmente auditáveis (Cohen, 1995). Embora não seja nativamente *multilabel*, empregamos novamente a estratégia um-contra-todos.

O **Random Forest** combina múltiplas árvores de decisão treinadas sobre subconjuntos aleatórios de instâncias e atributos (Breiman, 2001). Embora cada árvore individual possua uma estrutura simbólica e interpretável, a agregação de dezenas ou centenas de árvores torna a lógica global do modelo difícil de inspecionar, o que o afasta de uma abordagem puramente simbólica e o aproxima de um modelo de "caixa-preta". Neste trabalho, o Random Forest é incluído como um método de transição entre paradigmas simbólicos e estatísticos, alinhado ao objetivo mais amplo de comparar abordagens com diferentes níveis de interpretabilidade. Além disso, o método oferece suporte natural à classificação multiclasse, com extensões que permitem seu uso em cenários *multilabel*.

4.2 Configuração experimental

O pré-processamento é frequentemente apontado como uma limitação de abordagens clássicas em contraste com modelos profundos que operam sobre texto bruto. Para avaliar seu impacto, cada córpus é analisado em duas configurações: com e sem pré-processamento extensivo. Considerando o forte desbalanceamento observado na maioria dos córpus, sobretudo pela predominância da classe neutra, avaliamos adicionalmente um cenário em que todas as instâncias rotuladas como neutras foram excluídas do conjunto de dados. Sendo assim, os experimentos foram conduzidos sobre os córpus originais e também sobre subconjuntos contendo apenas amostras associadas a classes emocionais,

sem qualquer ocorrência da classe neutra.

Os córpus são divididos em 80% para treino e 20% para teste, com amostragem estratificada. Os experimentos são conduzidos de forma independente por córpus e por configuração. Quando aplicado, o pré-processamento extensivo inclui: (i) remoção de *hashtags* emocionais, evitando a trivialização da tarefa; (ii) remoção de *stop words*, acentos e caracteres especiais; (iii) normalização de *stocks*[3], URLs, valores monetários e porcentagens (da Silva et al., 2020); e (iv) lematização com *Stanza* (Qi et al., 2020).

Nos métodos supervisionados, os textos são representados por vetores TF-IDF (Manning et al., 2008), utilizando *scikit-learn* (Pedregosa et al., 2011). O RIPPER é aplicado diretamente sobre essa representação por meio da biblioteca *Wittgenstein* (Moscovitz, 2020). Árvores de decisão e *Random Forest* são treinados com *scikit-learn*, com seleção de hiperparâmetros via *Grid Search*. O espaço de busca inclui número de árvores (50, 100, 200, 500), profundidade máxima (nenhuma, 10, 20) e parâmetros mínimos de divisão (2, 5) e de folhas (1, 2).

5 Avaliação

Esta seção apresenta uma visão geral dos resultados experimentais, destacando tendências de desempenho entre os métodos avaliados e as diferenças observadas entre córpus e domínios. Também comparamos as abordagens sob a perspectiva do compromisso entre desempenho e interpretabilidade.

5.1 Resultados quantitativos

A Tabela 4 resume o desempenho dos métodos em cada córpus, utilizando *Medida-F macro* como métrica principal, por atribuir peso igual às classes e ser adequada a cenários desbalanceados (Sokolova and Lapalme, 2009). No córpus Domínio Livre sem pré-processamento, a anotação baseada em *hashtags* emocionais torna a tarefa trivial; sua remoção explica a queda acentuada de desempenho após o pré-processamento. De modo geral, observa-se um padrão estável de desempenho relativo: árvore de decisão e Random Forest obtêm os melhores resultados médios, seguidos por RIPPER, LexicoR e ZeroR. Essa ordenação se mantém na maioria dos córpus e configurações, com variações atribuídas ao domínio e ao esquema de anotação.

[3]Códigos de ações do mercado financeiro.

Embora o Random Forest generalize a árvore de decisão, esta última apresenta desempenho médio superior. Esse resultado é consistente com o desenho experimental: os modelos operam sobre representações esparsas (TF-IDF), com conjuntos de treinamento relativamente pequenos. Nessas condições, a combinação de múltiplas árvores com amostragem aleatória tende a diluir sinais lexicais fortes, enquanto uma única árvore determinística (por emoção) explora mais diretamente os atributos emocionalmente informativos. Esse resultado reforça que maior complexidade e custo computacional não implicam necessariamente melhor desempenho, especialmente quando o domínio não favorece a estrutura de modelos mais complexos específicos.

A Figura 1 apresenta o desempenho médio por córpus. A inclusão da classe neutra impacta negativamente os resultados ao intensificar o desbalanceamento e a ambiguidade entre classes. Apesar de resultados mistos nas árvores de decisão, o pré-processamento beneficia os métodos de forma geral, com ganhos mais consistentes nas abordagens simbólicas. Domínios em que as emoções são expressas por marcadores lexicais explícitos (Domínio Livre) ou vocabulário previsível (Notícias Curtas) apresentam melhor desempenho, enquanto domínios mais específicos e dependentes de contexto (DANTEStocks), exibem quedas generalizadas. Esses resultados indicam grande variação no desempenho absoluto entre domínios, mas estabilidade no desempenho relativo entre os métodos, reforçando a importância de avaliações comparativas multi-domínio.

5.2 Análise qualitativa

Do ponto de vista da interpretabilidade, observa-se um gradiente claro entre os métodos avaliados. O LexicoR e o RIPPER produzem resultados diretamente interpretáveis por meio de regras explícitas que associam termos ou combinações lexicais a emoções. Enquanto o desempenho do LexicoR depende fortemente da qualidade e cobertura do léxico, o que é um desafio diante da escassez desses recursos, o RIPPER surge como uma alternativa particularmente atraente.

As regras induzidas pelo RIPPER são simples e interpretáveis (Tabela 5): cada rótulo é definido por um conjunto de regras disjuntivas, bastando a satisfação de uma delas. Essas regras capturam termos emocionais (“incrível”, “ótimo”, “animado” e “empolgado”), modificadores de sentido (como

córpus	***ZeroR***	***LexicoR***	***Arv. Dec.***	***RIPPER***	**Rand. Forest**
DANTE_ComNeutro	0,0476	0,0921	**0,3120**	0,0950	0,1699
DANTE_ComNeutro_Preprocessado	0,0476	0,1269	**0,3152**	0,0715	0,1673
DANTE_SemNeutro	0,0689	0,0708	**0,3418**	0,0764	0,1887
DANTE_SemNeutro_Preprocessado	0,0689	0,0891	**0,3149**	0,0785	0,2110
DANTEConcordancia_ComNeutro	0,0966	0,1705	**0,2010**	0,1289	0,0961
DANTEConcordancia_ComNeutro_Preprocessado	0,0966	0,1172	**0,1711**	0,1191	0,0966
DANTEConcordancia_SemNeutro	0,0962	0,0000	0,1462	0,1197	**0,1855**
DANTEConcordancia_SemNeutro_Preprocessado	0,0962	0,0812	**0,2318**	0,1215	0,2150
DominioLivre	0,0807	0,0853	**0,9858**	0,9850	0,9777
DominioLivre_Preprocessado	0,0807	0,1299	**0,5022**	0,3613	0,4547
GoEmotionsBR_ComNeutro	0,0177	0,0422	**0,3141**	0,2506	0,1867
GoEmotionsBR_ComNeutro_Preprocessado	0,0177	0,0540	**0,2842**	0,2237	0,2073
GoEmotionsBR_SemNeutro	0,0088	0,0309	**0,3436**	0,2886	0,2087
GoEmotionsBR_SemNeutro_Preprocessado	0,0088	0,0442	**0,3236**	0,2827	0,2373
GoEmotionsBREkman_ComNeutro	0,0822	0,1809	**0,3959**	0,2007	0,2856
GoEmotionsBREkman_ComNeutro_Preprocessado	0,0822	0,2727	**0,3886**	0,2195	0,3427
GoEmotionsBREkman_SemNeutro	0,1213	0,1474	**0,4755**	0,2567	0,3830
GoEmotionsBREkman_SemNeutro_Preprocessado	0,1213	0,2631	**0,4777**	0,2788	0,4361
NoticiasCurtas_ComNeutro	0,0357	0,1041	**0,5083**	0,1881	0,4042
NoticiasCurtas_ComNeutro_Preprocessado	0,0357	0,1741	**0,5273**	0,2642	0,4698
NoticiasCurtas_SemNeutro	0,0476	0,1163	**0,4766**	0,2456	0,4441
NoticiasCurtas_SemNeutro_Preprocessado	0,0476	0,1501	**0,5289**	0,3352	0,5027
Media Geral	0,0639	0,1155	**0,3893**	0,2359	0,3123

Tabela 4: Resumo de desempenho (Medida-F macro) de todos os métodos de classificação por córpus.

negação) e expressões compostas frequentes (“bom trabalho”). Contudo, mudanças substanciais de domínio exigem nova indução. Em contraste, o LexicoR é estático e dispensa retreino, mas requer ampla cobertura e atualização contínua. Por não demandar recursos extras, como um léxico, o RIPPER é especialmente adequado a cenários com recursos limitados e alta exigência de interpretabilidade.

Emoção	***Regra***	
admiracao	[incrivel=V]	*OU*
	[otimo=V]	*OU*
	[bom=V *E* trabalho=V]	
empolgacao	[nao=F *E* parecer=F *E* estar=V *E* animado=V]	*OU*
	[nao=F *E* empolgado=V]	

Tabela 5: Regras fornecidas pelo RIPPER para o córpus GoEmotionsBR pré-processado (=V indica a presença e =F a ausência de cada termo, respectivamente).

Entre os métodos de melhor desempenho, a árvore de decisão apresenta interpretabilidade intermediária: embora seus caminhos de decisão sejam compreensíveis, o conhecimento encontra-se distribuído em múltiplos ramos, o que dificulta uma visão global do comportamento do modelo. A Figura 2, que representa parte da subárvore dos quatro primeiros níveis para a classe “medo” no córpus Domínio Livre pré-processado, mostra que o modelo identifica termos emocionais (“medo”, “assustador” e “terror”) e modificadores de sentido relevantes, como negação. Destaca-se a seleção do termo “estranho”, cuja relevância é coerente com o domínio do córpus, composto por publicações sobre o mercado de ações, em que variações inesperadas podem ser associadas ao medo dos acionistas. Essas evidências também ilustram o rápido crescimento do modelo, o que aumenta a complexidade interpretativa. O Random Forest, por sua vez, é substancialmente mais opaco, dificultando a inspeção direta de decisões individuais.

A análise qualitativa dos erros (Tabela 6) revela dois padrões principais nas falhas do método de melhor desempenho (árvores de decisão): (i) erros por omissão, nos quais um ou mais rótulos em exemplos *multilabel* não são previstos; e (ii) confusões em casos com maior nuance ou ambiguidade emocional. Em ambos, os erros estão associados a sinais emocionais fracos (linhas 4 e 5), textos curtos com pouca informação contextual ou ambíguos (linhas 2 e 4) e à sobreposição semântica entre emoções (linhas 1, 3, 4 e 5), evidenciando limites intrínsecos da tarefa.

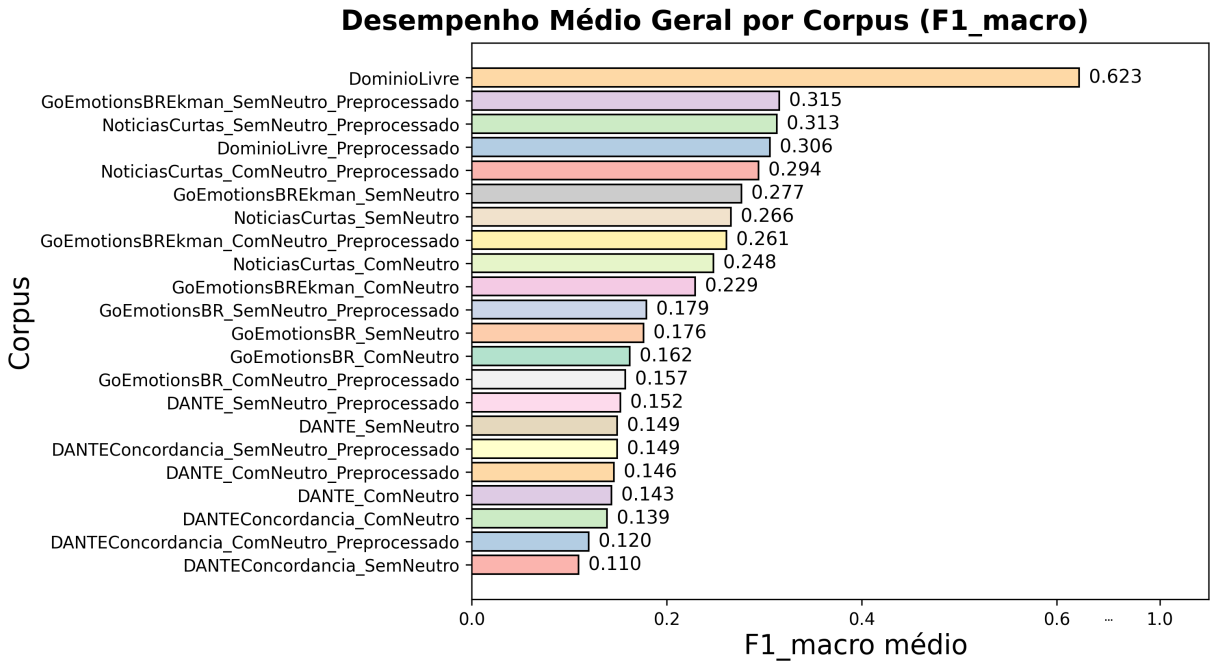


Figura 1: Desempenho (Medida-F macro) médio de cada córpus avaliado.

Texto	***Esperado***	***Previsto***
ela desrespeitou [NOME], não gosto dela	nojo, raiva	raiva
Parece deliciosamente entediante.	raiva	alegria
Sem contar que o vídeo completo faz parte da campanha, segundo o idiota do [NOME].	tristeza	raiva
Tenho certeza de que isso subiu para pelo menos 10 mil inscritos. Pode ser até mais alto agora!	alegria	surpresa
Filhotes de caracal nascem em cativeiro: Espécie está ameaçada de extinção em Israel.	surpresa	alegria

Tabela 6: Exemplos de erros da árvore de decisão. O marcador [NOME] foi introduzido apenas neste artigo para fins de reprodução, com o objetivo de evitar a identificação e associação com as figuras públicas citadas.

6 Impressões e discussões teóricas

Os córpus analisados abrangem desde textos formais, como manchetes jornalísticas, até conteúdos informais de redes sociais, permitindo avaliar o efeito do estilo textual na detecção de emoções. Textos mais estruturados e emocionalmente ricos, como obras literárias, seriam relevantes, mas não encontramos recursos públicos em português com essas características.

Os resultados mostram que métodos simbólicos são mais eficazes em domínios com emoções explícitas, vocabulário previsível e menor ambiguidade. Nesses cenários, maior complexidade não implica melhor desempenho, como no caso das árvores de decisão frente ao Random Forest. Observa-se, assim, um compromisso entre desempenho e interpretabilidade: modelos mais complexos tendem a obter melhores médias, enquanto métodos baseados em regras oferecem resultados mais transparentes, fundamentais em aplicações sensíveis.

A análise qualitativa dos erros evidencia limites intrínsecos da tarefa, sobretudo em cenários *multi-label* e com sinais emocionais sutis ou ambíguos. Os córpus avaliados adotam modelos de emoção com pressupostos distintos, que vão de abordagens baseadas em evidências fisiológicas inatas (Ekman) a esquemas derivados de *crowdsourcing* (GoEmotions) e modelos que assumem exclusividade mútua entre estados emocionais (Roda de Plutchik), evidenciando a dificuldade de definir emoções de forma consensual. O domínio e o esquema de anotação influenciam fortemente os resultados, e córpus desbalanceados ou com expressão emocional sutil permanecem desafiadores.

Observa-se também forte desbalanceamento de polaridade: o GoEmotions-BR é o modelo mais granular, com múltiplas emoções positivas, enquanto os demais córpus assumem, em geral, apenas a classe “alegria”. Modelos mais granulares ampliam a ambiguidade e a subjetividade, dificultando distinguir erros de anotação de limitações do classificador. A classe neutra é particularmente indefinida, frequentemente associada a expressões

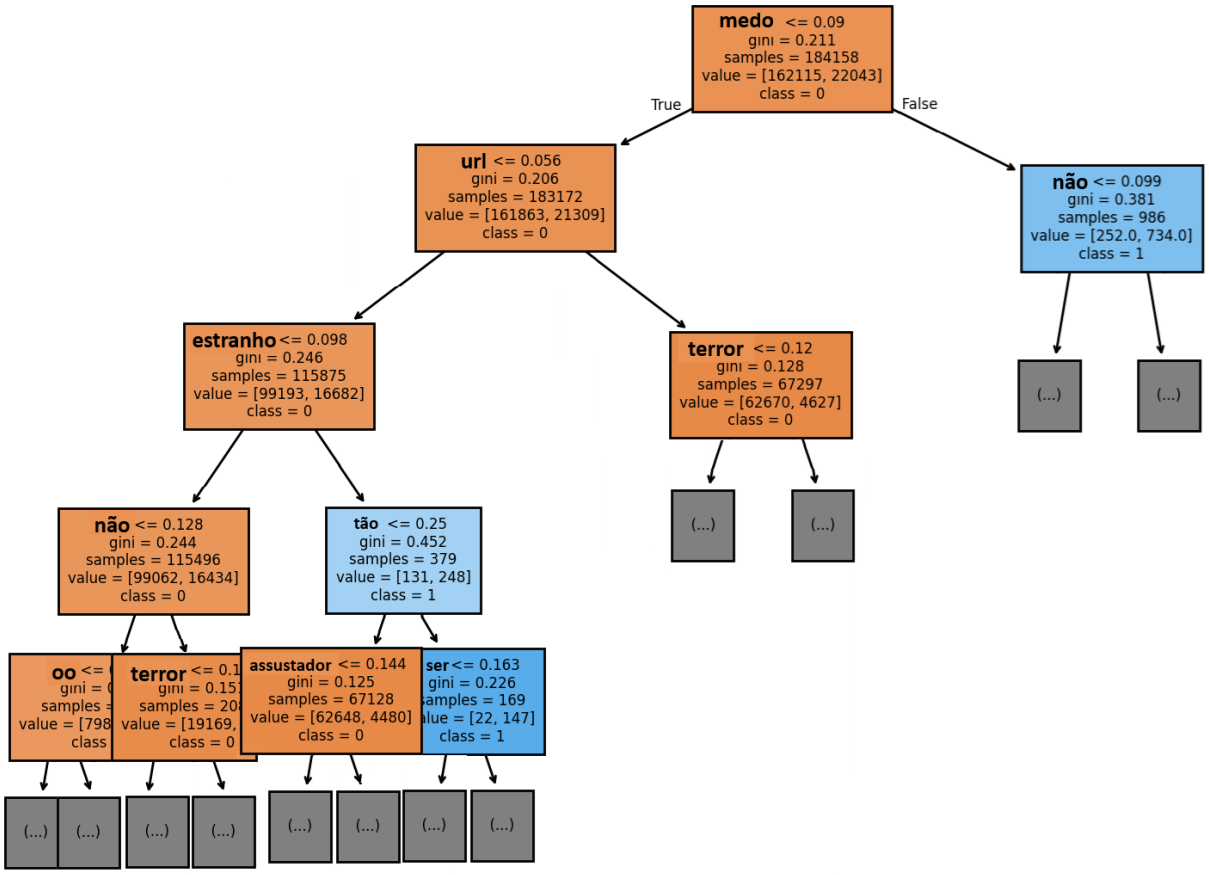


Figura 2: Árvore de decisão da classe “medo” para o córpus Domínio Livre pré-processado. A cor de cada nó indica a classe majoritária predita pela subárvore enraizada nesse nó. Neste caso, as classes correspondem à presença (azul) ou à ausência (laranja) do rótulo “medo”. A intensidade da cor é proporcional à pureza do nó, definida pela proporção de amostras da classe majoritária associadas a esse nó.

fracas ou ambíguas, o que intensifica a confusão entre classes. Em síntese, não há um método universalmente superior; a escolha depende do contexto, do modelo de emoção e dos requisitos de interpretabilidade.

Como trabalhos futuros, destacam-se abordagens híbridas simbólico-neurais e a indução automática de regras a partir de representações linguísticas mais ricas que o TF-IDF, como *word embeddings* aprendidos por modelos neurais, visando conciliar desempenho, robustez e interpretabilidade. Pretendemos também avançar no estudo mais amplo do qual este trabalho faz parte, ampliando a comparação entre paradigmas de classificação, modelos de emoção e domínios, para analisar sistematicamente suas vantagens e limitações no contexto do português.

Limitações

As principais limitações incluem sensibilidade ao domínio, dependência de vocabulário emocional explícito e menor transparência nos modelos de melhor desempenho. No método LéxicoR, o desempenho depende diretamente da qualidade e cobertura dos recursos, ainda escassos para a língua portuguesa.

Considerações éticas

Os classificadores de emoções aqui apresentados baseiam-se em modelos teóricos com limitações inerentes, dados os desafios de obter consenso científico sobre a definição e a delimitação das emoções humanas. Assim, suas predições não devem ser tratadas como evidências objetivas nem utilizadas em contextos sensíveis ou para tomada de decisão automatizada, devendo restringir-se a uso auxiliar e informativo, sob responsabilidade e supervisão humana.

Ademais, essas ferramentas suscitam riscos éticos relevantes, incluindo a possibilidade da informação ser utilizada para manipulação ou indução de estados emocionais para fins comerciais e/ou

políticos. Questões adicionais envolvem vieses nos dados e nos modelos, falta de transparência e riscos à privacidade e ao consentimento, reforçando a necessidade de cautela quanto à interpretação e ao uso desses sistemas.

Agradecimentos

Agradecemos aos autores dos recursos utilizados por disponibilizarem os dados para pesquisa e contribuírem no avanço da área para o português.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES). Este projeto também foi apoiado pelo Ministério da Ciência, Tecnologia e Inovações, com ecursos da Lei N. 8.248, de 23 de outubro de 1991, no âmbito do PPI-Softex, coordenado pela Softex e publicado como Residência em TIC 13, DOU 01245.010222/2022-44.

Referências

Amer F. A. H. Alnuaimi and Tasnim H. K. Albaldawi. 2024. An overview of machine learning classification techniques. *BIO Web of Conferences*, 97:00133.

Leo Breiman. 2001. Random Forests. *Machine Learning*, 45(1):5–32.

William W. Cohen. 1995. Fast Effective Rule Induction. In Armand Prieditis and Stuart Russell, editors, *Machine Learning Proceedings 1995*, pages 115–123. Morgan Kaufmann, San Francisco (CA).

Diogo Cortiz, Jefferson Silva, Newton Calegari, Ana Freitas, Ana Soares, Carolina Botelho, Gabriel Rêgo, Waldir Sampaio, and Paulo Boggio. 2021. A Weakly Supervised Dataset of Fine-Grained Emotions in Portuguese. In *Anais do XIII Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 73–81, Porto Alegre, RS, Brasil. SBC.

Fernando J. Vieira da Silva. 2020. *Cross-domain emotion detection in tweets*. Tese de doutorado, Universidade Estadual de Campinas, Instituto de Computação, Brasil.

Fernando J. Vieira da Silva, Norton T. Roman, and Ariadne M. B. R. Carvalho. 2020. Stock market tweets annotated with emotions. *Corpora*, 15(3):343–354.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. GoEmotions: A Dataset of Fine-Grained Emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Mariza Miola Dosciatti, Lohann Paterno Coutinho Ferreira, and Emerson Cabrera Paraiso. 2013. Identificando emoções em textos em português do brasil usando máquina de vetores de suporte em solução multiclasse. *X ENIAC: Encontro Nacional de Inteligência Artificial e Computacional. Fortaleza, Brasil*.

Luís Carlos Fernandes Duarte. 2019. Reconhecimento automático de emoções em texto com recurso a emojis. Dissertação de mestrado, Universidade de Coimbra, Portugal.

Paul Ekman and Dacher Keltner. 1997. Universal facial expressions of emotion: An old controversy and new findings. In Ullica C. Segerstråle and Peter Molnár, editors, *Nonverbal Communication: Where Nature Meets Culture*, pages 27–46. Lawrence Erlbaum Associates, Mahwah, NJ.

Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2018. A Survey of Methods for Explaining Black Box Models. *ACM Comput. Surv.*, 51(5).

Luiz Otávio Alves Hammes and Larissa Astrogildo de Freitas. 2021. Utilizando BERTimbau para a Classificação de Emoções em Português. In *Anais do XIII Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana (STIL)*, pages 56–63, Porto Alegre, RS, Brasil. SBC.

C. D. Manning, P. Raghavan, and H. Schütze. 2008. *Introduction to Information Retrieval*. An Introduction to Information Retrieval. Cambridge University Press.

Abdullah Al Maruf, Fahima Khanam, Md. Mahmudul Haque, Zakaria Masud Jiyad, M. F. Mridha, and Zeyar Aung. 2024. Challenges and Opportunities of Text-Based Emotion Detection: A Survey. *IEEE Access*, 12:18416–18450.

Saif M. Mohammad. 2016. 9 - Sentiment Analysis: Detecting Valence, Emotions, and Other Affectual States from Text. In Herbert L. Meiselman, editor, *Emotion Measurement*, pages 201–237. Woodhead Publishing.

Ilan Moscovitz. 2020. wittgenstein: Ruleset covering algorithms for transparent machine learning (RIPPER). `https://github.com/imoscovitz/wittgenstein`. Acesso em: 2026-01-29.

Sreerama K. Murthy. 1998. Automatic construction of decision trees from data: A multi-disciplinary survey. *Data Mining and Knowledge Discovery*, 2(4):345–389.

Francisco Bráulio Oliveira and Jaime Simão Sichman. 2024. Portuguese Emotion Detection Model Using BERTimbau Applied to COVID-19 News and Replies. In *Anais da XXXIV Brazilian Conference on Intelligent Systems (BRACIS)*, pages 265–280, Porto Alegre, RS, Brasil. SBC.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12(85):2825–2830.

Robert Plutchik. 2001. The Nature of Emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *American Scientist*, 89(4):344–350.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108. Association for Computational Linguistics.

Barbara Ramos. 2021. Descrição de uma metodologia desenvolvida para revisão de um léxico de palavras de emoção. In *Anais do XIII Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 389–397, Porto Alegre, RS, Brasil. SBC.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 1135–1144, New York, NY, USA. Association for Computing Machinery.

Allisfrank dos Santos. 2019. Análise de sentimento multiclasse: uma abordagem com o uso de aprendizado de máquina. Dissertação de mestrado, Universidade Federal de São Carlos, Brasil.

Dibyendu Seal, Uttam K. Roy, and Rohini Basak. 2020. Sentence-level emotion detection from text based on semantic rules. In *Information and Communication Technology for Sustainable Development: Proceedings of ICT4SD 2018*, pages 423–430. Springer Singapore.

Marina Sokolova and Guy Lapalme. 2009. A systematic analysis of performance measures for classification tasks. *Information Processing Management*, 45(4):427–437.

Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. 2020. BERTimbau: Pretrained BERT Models for Brazilian Portuguese. In *Intelligent Systems: 9th Brazilian Conference (BRACIS 2020)*, pages 403–417. Springer International Publishing.

A Implementação do LexicoR

Com o objetivo de facilitar a compreensão e a reprodutibilidade do trabalho, este apêndice detalha o funcionamento do algoritmo LexicoR, proposto na Seção 4.1. O pseudo-código para a classificação de uma amostra é apresentado no Algoritmo 1.

O léxico de emoções (Emocionário), que associa palavras a emoções, é lido e a associação é mantida em uma estrutura de dicionário. São definidas as listas de modificadores semânticos de negação, intensificação e redução, apresentadas na Tabela 7.

Tipo de Modificador	***Lista de Palavras***
Negação	"jamais", "nada", "nem", "nenhum", "ninguém", "ninguem", "nunca", "não", "nao", "tampouco", "sem", "sequer"
Intensificação	"mais", "muito", "demais", "completamente", "absolutamente", "totalmente", "definitivamente", "extremamente", "frequentemente", "bastante", "super", "altamente", "incrivelmente", "intensamente", "profundamente", "terrivelmente", "tão", "tao", "enormemente"
Redução	"pouco", "quase", "menos", "apenas", "levemente", "ligeiramente", "parcialmente", "raramente", "ocasionalmente", "moderadamente"

Tabela 7: Listas de palavras associadas a cada tipo de modificação.

Para classificar uma amostra, cada *token* da amostra é analisado individualmente. Se o *token* também estiver presente no Emocionário, então essa palavra representa alguma das emoções mapeadas pelo léxico emocional. Nesse caso, o algoritmo incrementa a pontuação da classe (emoção) em questão.

Para definir o valor do incremento, uma janela com as quatro palavras que apareceram antes do *token* atual e com as quatro palavras que aparecem depois do *token* é considerada. Se não houver modificadores de negação, intensificação ou redução na janela considerada, a pontuação da classe é incrementada em 1. Caso haja algum modificador de negação, o incremento da pontuação é -3. Os modificadores de intensificação e redução alteram o incremento para 3 e 1/3, respectivamente. Se um modificador de negação aparecer juntamente de algum dos outros modificadores, o incremento é o inverso: 3 para 1/3 e vice-versa.

Depois de repetir este processo para cada *token* da amostra, o algoritmo escolhe e prevê as classes com a maior pontuação para a amostra. Caso a amostra não possua nenhum marcador de emoção

do Emocionário, a pontuação de todas as classes será zero e a classe neutra será atribuída à amostra.

Algoritmo 1 Pseudo-código do LexicoR

Entrada: Amostra a ser classificada a
Entrada: Léxico emocional $emoc$
Saída: Classes de emoção preditas c

$neg \leftarrow$ conjunto de modificadores de negação
$inten \leftarrow$ conjunto de modificadores de intensificação
$reduc \leftarrow$ conjunto de modificadores de redução

$scores \leftarrow$ vetor com o valor 0 para todas as classes de emoções em $emoc$
para cada *token* p de a **do**
    se $p \in emoc$ **então**
        $label \leftarrow$ emoção associada a p no léxico $emoc$
        $ctx \leftarrow$ 4 palavras anteriores e 4 palavras posteriores a p em a
        se ctx contém alguma palavra de $inten$ **então**
            se ctx contém alguma palavra de neg **então**
                $peso \leftarrow 1/3$
            senão
                $peso \leftarrow 3$
            fim se
        senão se ctx contém alguma palavra de $reduc$ **então**
            se ctx contém alguma palavra de neg **então**
                $peso \leftarrow 3$
            senão
                $peso \leftarrow 1/3$
            fim se
        senão se ctx contém alguma palavra de neg **então**
            $peso \leftarrow -3$
        senão
            $peso \leftarrow 1$
        fim se
        $scores[label] \leftarrow scores[label] + peso$
    fim se
fim para

se todos os valores de $score$ forem 0 **então**
    retorne "neutro"
fim se

$maior \leftarrow max(\{score\})$
retorne todas as classes $label$ tal que $score[label] = maior$

Robustness and Diversity Evaluation on ProsSegue-ML: a Free Prosodic Segmentation Tool for Brazilian Portuguese

Giovana Meloni Craveiro
ICMC-USP, BRAZIL
giovana.meloni.craveiro@alumni.usp.br

Sandra Maria Aluísio
ICMC-USP, BRAZIL
sandra@icmc.usp.br

Abstract

Prosodic segmentation is the task of dividing a sound unit into smaller units, which can be distinguished between units with a completed idea, marked by TBs, and non-autonomous units, marked by NTBs. Enhancing the performance of ASR and TTs systems is a useful task, and it remains relevant for Brazilian Portuguese due to the diversity of conditions and speaker-related factors that influence its performance. Here, we explore a low-impact, open-source approach based on a Random Forest classifier and a set of features that include fundamental frequency, speech rate, pauses, and energy (Craveiro et al., 2025). We perform a robustness evaluation of the referred ML model, modifying a few conditions on its training, comparing its performance when tested in other datasets, and comparing its results with those of other studies using the same data samples. We experiment with augmenting the training dataset and evaluating how the bias of speaker profile aspects is affected when the size and diversity of the training set are changed. Although we don't achieve statistically significant values in the bias evaluation, we observe that inequalities grow as the training dataset is expanded with a much larger, but less diverse sample of data.

1 Introduction

Information in spoken language is conveyed not only through lexical items, but also through a range of non-segmental features, commonly referred to as prosodic cues, including pitch, intensity, speech rate, rhythm, and timbre. Speech segments delimited by such prosodic cues are capable of expressing coherent messages and fulfilling a variety of linguistic functions, which are realized through different types of utterance (e.g., imperative, interrogative, assertive, or exclamatory). These prosodically delimited segments are typically referred to as intonational phrases or intonation units (IUs). Although IUs are difficult to define precisely, they are generally characterized by the presence of a well-defined (i.e., "single") pitch contour (Biron et al., 2021).

There are studies, such as (Mello et al., 2012; Santos et al., 2022) that distinguish between terminal break (TB) units, which mark complete sequences, that is, they communicate the conclusion of an idea, constituting the smallest pragmatically autonomous unit of speech, and non-terminal break (NTB) units, which signal a non-autonomous unit, whose information is not completed within the same unit. The identification of these boundaries is based on prosodic cues, such as variations in fundamental frequency (F0), segment duration, and the presence of pauses, in addition to inspection of the acoustic signal. In this work, we will follow the same distinction, but we will focus only on terminal boundaries, as we chose a method that does not segment NTBs.

Automatic detection of prosodic boundaries in natural language speech has been extensively investigated in the speech processing literature (Wightman and Ostendorf, 1991; Ananthakrishnan and Narayanan, 2008; Huang et al., 2008; Jeon and Liu, 2009; Kocharov et al., 2017; Biron et al., 2021). Despite substantial progress, this task remains challenging due to the numerous sources of variability inherent in speech signals. These sources include speaker-related factors (e.g., age, gender, and dialectal variation), recording conditions (such as microphone type, room acoustics, and background noise), and production style, ranging from spontaneous to read speech, which can be understood as points along a continuum between unplanned and planned speech production. Furthermore, machine learning methods are prone to biases (Broussard, 2018; Buolamwini and Gebru, 2018; Ruback et al., 2022). In the prosodic segmentation scenario, (Craveiro and Galdino, 2025) argues that it is imperative to select a corpus that is diverse in terms

Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)
Vol. 2, pages 170–180
April 13-16, 2026

of accent, gender, age, and educational level.

Accurate prosodic boundary detection has direct implications for both automatic speech recognition (ASR) and text-to-speech (TTS) systems. In ASR, training models for speech excerpts segmented according to IUs have been shown to reduce syllable, character, and word-level error rates (Chen and Hasegawa-Johnson, 2004; Lin et al., 2019). In TTS systems, appropriate modeling of prosodic phenomena, such as pause duration, naturally used by human speakers, contributes to improved speech intelligibility and more effective transmission of meaning (Liu et al., 2022). Consequently, effective automatic identification of prosodic boundaries is expected to (i) facilitate linguistic analysis of spontaneous speech, (ii) support the creation of more informative datasets for ASR and TTS training, and (iii) improve the performance of speech-related applications operating on spontaneous speech (Galdino et al., 2026b,a).

Approaches to automatic prosodic boundary detection range from rule-based or heuristic systems, e.g. (Biron et al., 2021), to supervised machine learning models that integrate lexical and syntactic information with acoustic features, such as (Kocharov et al., 2017). The set of acoustic features differs in each study, but many of them include features related to pauses, speech rate, amplitude, and fundamental frequency (Kocharov et al., 2017; Raso et al., 2020; Biron et al., 2021). Such methods have been predominantly applied to scripted speech, where syntactic and prosodic structures tend to align, and disfluencies are relatively rare. More recently, Roll et al. (2023) proposed fine-tuning Whisper (Radford et al., 2023), a pretrained end-to-end ASR model, to segment spontaneous speech into intonation units, achieving strong performance.

Research on automatic prosodic boundary detection for Brazilian Portuguese has been conducted mainly by the speech processing group of the Federal University of Minas Gerais (Teixeira et al., 2018; Raso et al., 2020; Teixeira, 2022). However, no segmentation tool was made publicly available, hindering the easy application of the method to various studies in this language. In an effort to promote the public availability of tools, (Craveiro et al., 2024, 2025; Galdino et al., 2026b) have made open-source resources and models available for the task of prosodic segmentation of Brazilian Portuguese.

However, the low-impact open-source model from (Craveiro et al., 2025), which is the most recent approach and was trained and tested on a diverse dataset (balanced in gender and relatively diverse in terms of accents, ages, and educational levels), MuPe-Diversidades (Craveiro and Galdino, 2025), is no longer replicable due to an update in the version of the forced phonetic aligner, UFPAlign (Batista et al., 2022), used in the study. In addition to this problem, the machine learning model they evaluated includes a feature based on the difference between the F0 average of a syllable and the F0 average of the TB it belongs to, implying that it requires prior annotation of TBs. Furthermore, the authors removed all the questions uttered by the interviewers, focusing only on the respondents' answers. The authors also performed a bias evaluation on their model, and their results suggest a biased performance depending on the profile of the speaker, but the values obtained did not achieve statistical significance.

This work starts from the scenario above, which impacts the public availability of functional models for the task of prosodic segmentation, and aims to answer 4 research questions, listed below. All the models trained here and the resources used to answer the questions are available on the website: `https://github.com/nilc-nlp/ProsSegue`

1. What is the impact of using the current version of UFPAlign instead of UFPAlign's prior version? What is the impact of removing the feature that requires a previous annotation of TBs, using only 8 features? Also, what is the impact of modifying such a feature by using the F0 average of units separated by silent pauses, instead of the F0 average of units separated by TBs? And, what is the impact of training the model with and without the interviewers' speech? (see Section 4.1)

2. Considering that the test set from MuPe-Diversidades is comprised of speech from the same speakers that are present in the training set, it is especially relevant to assess the robustness of the model. Thus, what is the performance of the best model resulting from Question 1 when tested in different corpora, such as NURC-CM and samples from C-oral I and II? (see Section 4.2)

3. Is the model equally effective for diverse speaker profiles (in terms of gender, age, region of birth, and educational levels) if trained

in a less diverse but larger dataset (NURC-SP MC)? (see Section 4.3)

4. How is bias affected when augmenting a diverse dataset with a significantly larger but less diverse sample of data (MuPe-Diversidades + NURC-SP Minimum Corpus)? (see Section 4.3)

2 Related Work

Various approaches (rule-based, traditional machine learning, and deep learning) have been proposed to address the challenge of automatic prosodic segmentation (see Table 1 for studies that focus on Portuguese and English).

In (Kocharov et al., 2017), intonational units were predicted by combining syntactic and acoustic features using a Random Forest classifier. Applied to American English (Boston University Radio Speech Corpus), the study reported an F1 measure of 76% using prepared speech; for Russian, the language for which the method was originally proposed, it obtained an F1 equal to 91% in the Corpus of Professionally Read Speech. (Biron et al., 2021) used heuristics based on pause duration and speech rate discontinuities to detect prosodic boundaries in spontaneous speech from American English (Santa Barbara Corpus of Spoken American English - SBCSAE). With Montreal Forced Aligner and evaluation in Praat, the study indicated a performance of 66% on the F1 measure. By fine-tuning the Whisper model (Radford et al., 2023), (Roll et al., 2023) proposed a method (named PSST) that integrates prosodic and lexical-syntactic information for the segmentation of spontaneous speech, and functions also as a transcription tool. It achieved 87% F1 measure for American English (SBCSAE) and 73% F1 measure for British English (Intonational Variation in English (IViE) corpus - urban dialects of English spoken in the British Isles). The authors suggest that at least some of the success of PSST is due to the interaction of acoustic and lexico-syntactic information, which arises due to its integration of IU boundary detection with STT transcription.

For European Portuguese, (Hoi et al., 2022) detected boundaries through spectrograms and a convolutional neural network, using prepared speech. The technique achieved 95.6% accuracy and works for any language, but it is based solely on pauses, excluding the possibility of identifying units that do not end with silences.

(Raso et al., 2020) developed a linear discriminant analysis (LDA) classifier applied to spontaneous speech in Brazilian Portuguese, based on acoustic parameters. They use samples from C-ORAL BRASIL I and II, with prosodic boundaries annotated by experts. 111 phonetic-acoustic features were extracted, via Praat script, from the speech signal corresponding to all V-V units in windows centered on the boundaries between phonological words. The extracted features comprised 5 groups of measures: 1) Speech rate and rhythm; 2) Normalized duration; 3) Fundamental frequency; 4) Intensity; 5) Silent pause (presence and duration). Positions at which at least 50% of the annotators indicated a boundary of the same type were considered a boundary. Several models were trained to identify terminal boundaries (TBs) and non-terminal boundaries (NTBs): (i) the TB-b1 model, with pause and F0 as main parameters, was trained on Sample I (balanced), and the test on Sample II had an accuracy of 76.3% for TBs; (ii) the TB-b2 model was trained on Sample II (balanced), and the test on Sample I had an accuracy of **80.8%** for TBs; Features related to pauses and F0 were the main features associated with the identification of terminal boundaries. The best values of accuracy (in bold above) are in Table 1, for TB and NTB models. For spontaneous speech in BP, there is also the method by (Craveiro et al., 2024), which detected prosodic boundaries using the forced phonetic aligner UFPAlign (Batista et al., 2022) and the same heuristics as (Biron et al., 2021). The results indicated an F1 measure of 31%, using a 5-hour excerpt from the NURC-SP Minimal Corpus (MC), which reflects the linguistic variety of São Paulo. (Craveiro et al., 2025), inspired by the work of (Ananthakrishnan and Narayanan, 2008), used nine acoustic-prosodic features to train a Random Forest classifier, and reported binary and macro F1 measures of 55% and 77%, respectively, in the Mupe-Diversidades corpus (speech from 17 Brazilian states).

3 Methodology

3.1 Datasets

This work uses three datasets of spontaneous speech in Brazilian Portuguese that already contained annotation of prosodic segmentation: MuPe-Diversidades, NURC-SP Minimum Corpus, and samples extracted from C-ORAL BRASIL I and II.

MuPe-Diversidades is described in (Craveiro and

Table 1: Summary of prosodic segmentation research on prepared and spontaneous speech

Source	Language	Corpus	F1 Score/Accuracy	Open code?	Speech
Kocharov et al. (2017)	EN-US*	BURSC (∼10hs)	76%/86.5%	No	prepared
Biron et al. (2021)	EN-US*	SBCSAE (∼20hs)	66%/—	No	spontaneous
Roll et al. (2023)	EN-US EN-GB	SBCSAE (∼20hs) IViE (∼36hs)	87%/96% (SBC) 73%/93% (IViE)	open code	spontaneous
Hoi et al. (2022)	PT-PT*	RTP** (∼33hs)	—/95.6%	No	prepared
Raso et al. (2020)	PT-BR*	C-ORAL BRASIL I and II (∼17min) TB boundary	—/80.8%	No	spontaneous
Raso et al. (2020)	PT-BR	C-ORAL BRASIL I and II (∼17min) NTB boundary	—/75.6%	No	spontaneous
Craveiro et al. (2024)	PT-BR	Part of the NURC-SP MC (∼5hrs)	31%/—	open code	spontaneous
Craveiro et al. (2025)	PT-BR	MuPe-Diversidades (2h30min)	55%/97%	open code	spontaneous

*"EN-US" stands for American English, "EN-GB" for British English, PT-PT for European Portuguese, PT-BR for Brazilian Portuguese. **https://www.rtp.pt/

Galdino, 2025); it contains around 2.5 hours of speech extracted from life interviews of 30 people with diverse speaker profiles. The speakers were born in different cities from one of the 17 states comprised in the dataset: Alagoas, Bahia, Ceará, Paraíba, Pernambuco, Piauí, Sergipe, Pará, Rondônia, Goiás, Mato Grosso do Sul, Espírito Santo, Minas Gerais, Rio de Janeiro, São Paulo, Paraná, and Rio Grande do Sul. Each state present in the dataset is represented by 1 or 2 speakers, with excerpts of 10 or 5 minutes of speech, respectively. The corpus is also balanced in gender and diverse in age (20 to 91 years old) and educational level (no education, incomplete elementary school, complete elementary school, technical education, incomplete bachelor's degree, complete bachelor's degree, and master's degree).

Minimum Corpus is a subset of NURC-SP, composed of 21 audio files, with six formal lectures (EF), six dialogues between two informants (D2), and nine dialogues between one informant and one interviewer (DID). All of its speakers have superior education and are from the capital of São Paulo. Women are represented in 11 of the audios, and men are represented in 10. The speakers were categorized in age groups: group I: 25–35 years old, group II:36–55 years old, and group III:56 to 85 years old. The Minimum Corpus contains speech of 7 people from group I, 9 people from group II, and 5 people from group III. Since the recordings were made in the 1970s, the quality of the audios was also categorized, either as positive (good, very good, audible, clear), negative (low, very low, bass, noisy), or mixed evaluation (Santos et al., 2022). Since a few excerpts were removed due to failure of forced alignment, and one file was separated for the test set, the training set totals 17h35min19s,

C-ORAL BRASIL I and II are corpora of spontaneous speech in Brazilian Portuguese. C-ORAL I is entirely dedicated to informal speech and comprises 139 informal speech texts, and 21:08:52 hours of recording, distributed into family/private (80%) and public (20%) context. It is quite balanced in terms of speakers' gender, age, and school level (Raso and Mello, 2012). C-ORAL BRASIL II is dedicated to formal speech, comprising also a media and a telephonic corpus (Bossaglia and de Almeida Ferrari, 2019; Mello et al.). The prosodically segmented samples consist of fourteen approximately 1.5-minute excerpts of monologic male speech. Seven excerpts are drawn from C-ORAL BRASIL I (hereafter, Sample I), and seven from C-ORAL BRASIL II (Sample II), which include formal and media speech in natural contexts. The speakers represent the cities of Minas Gerais, Rio de Janeiro, Pará, São Paulo, and Santa Catarina. Age and education of these specific speakers were not disclosed. The total duration of the annotated corpus is approximately 17 minutes (Teixeira, 2022).

3.2 Models

The method that was chosen for the segmentation evaluation in this paper was reported in (Craveiro et al., 2025). It is a low-cost, low-impact approach based on a Random Forest classifier, trained with a diverse corpus, which is automatically phonetically aligned to identify the initial and final timestamps of each phone, syllable, and word. It covers solely TBs and considers 9 features at the syllable level, with the following order of importance: pause duration, energy range, difference between maximum and average energy, F0 range, nucleus vowel duration, difference between maximum and average F0, difference between minimum and average F0, difference between minimum and average energy, and difference between average F0 of the syllable and average F0 of the TB unit (f0_avgutt_diff). We assess the impact of performing a few modifications (detailed in Section 4.1) to the original model, generating a model we call MuDi.

We then perform further evaluations with MuDi

to analyze its performance on different corpora and compare its results with two other studies, using the same data samples (see Section 4.2). Finally, we experiment with expanding the training set with Corpus Minimum data. We train a model exclusively on 19 files from Minimum Corpus, excluding SP DID 234, which is separated as the test set, and name it MC. We evaluate MC's effectiveness across age, gender, educational level, and region of birth. We compare the performance of both models, analyzing whether there was any increase or decrease in biases caused by the profile of the speakers present in the training set. We also perform those evaluations on a model trained on a data sample composed of MuPe-Diversidades combined with NURC-MC, which we named MC-MuDi.

3.3 Evaluation

The results are calculated considering false positives, false negatives, true negatives, and true positives. False positives (FP) occur when the method falsely indicates a boundary. False negatives (FN) occur when the model does not identify a boundary that existed in the reference annotation. True positives (TP) occur when the model correctly identifies boundaries, and true negatives (TN) occur when the model correctly indicates a no-boundary position in places where there are no boundaries in the reference annotation. Each study uses a slightly different set of metrics, including a few of the following: accuracy, specificity, sensitivity/recall, precision, SER, macro F1 score, and binary F1 score. The two types of F1 score are differentiated according to the values considered. While the binary F1 score considers the existence of only one class: terminal breaks (TBs), the macro F1 score considers an average of the results of the class TB and the results of a secondary class that considers every position where there are no boundaries (NB). Such a category may not exist, according to the approach. (Craveiro et al., 2024), for instance, only has a class of type TB, but (Craveiro et al., 2025) considers the end of each syllable as a possible position for a boundary, so there are two classes: TBs and NBs.

4 Results and Discussion

4.1 Impact evaluation on updating the ProsSegue-ML model

This section details the results obtained when assessing the impact of making a few changes to the circumstances in which model ProsSegue-ML was trained. Table 2 compares different versions of models similar to the ProsSegue-ML model, but comprising one or more of the changes described below. The first line of the table indicates the model used in (Craveiro et al., 2025) and its last line indicates MuDi, the model we use in our further experiments.

ProsSegue-ML model relies on UFPAlign, a forced phonetic aligner developed for Brazilian Portuguese, which was updated in June of 2025. The update included changing its phonetic transcription and syllabification routines, which were independent from each other and implied a problematic procedure to align them, to a single routine that relies on a many-to-many (m2m) aligner [1] (git, 2025). Thus, the former UFPAlign version, which was used in (Craveiro et al., 2025), became obsolete. Here, we measure the impact of using this new version of UFPAlign to align the dataset (see line 2 of Table 2). There is a decrease of 1% in macro F1, which is acceptable since UFPAlign's former version is now obsolete.

Additionally, one of the 9 features used by (Craveiro et al., 2025), the difference between the average F0 of a syllable and the average F0 of the TB (f0_avg_utt_diff), requires a previous annotation of TBs, limiting the usability of the method to annotated datasets. Thus, we experiment with modifying how f0_avg_utt_diff is calculated by relying on the average F0 of units separated by pauses, instead of the F0 average of TBs. We name this altered feature f0_avg_utt_diff_2 and measure the impact of using it instead of f0_avg_utt_diff (line 3 of Table 2). We also measured the impact of simply removing such a feature (line 4 of Table 2). The table indicates that both of those experiments obtained a macro F1 of 77%, which is 1% higher than the macro F1 of the original set of features when the current version of UFPAlign is also used. We use 8 features with MuDi since it is simpler and actually yielded slightly better results (0.7711) than using f0_avg_utt_diff_2 (0.7698). It is hard to understand why this difference in performance occurred without a qualitative analysis of the segmentation, so, in future work, we intend to perform an analysis of the errors, searching for differences in alignment and segmentation among the versions of the classifier, as well as errors that occurred in each one.

[1] https://github.com/letter-to-phoneme/m2m-aligner

Table 2: Table comparing the performance of different versions of the model trained in MuPe-Diversidades. It shows the impact of changing UFPAlign's version, the set of features, and training audios (including and excluding interviewers' speech, which contains several questions). v1 = f0_avg_utt_diff and v2 = f0_avg_utt_diff_2.

Model	Features	UFPAlign	Questions included ?	macro F1
Craveiro et al. 2025 model	9 (v1)	obsolete version	no	0.77
UFPAlign evaluation	9 (v1)	current version	no	0.76
Features evaluation 1	9 (v2)	current version	no	0.77
Features evaluation 2	8	current version	no	0.77
Evaluation 4	8	obsolete version	no	0.77
Evaluation 5	9 (v2)	current version	yes	0.75
MuDi	8	current version	yes	0.75

Furthermore, (Craveiro et al., 2025) uses the corpus MuPe-Diversidades, which is composed of a series of excerpts of interviews. To train ProsSegue-ML, they removed the speech of interviewers in order to preserve the balance across speaker profiles [2]. However, all the questions contained in the speech of the interviewers were, therefore, removed. Thus, we also measure the impact of training a model without removing the speech of interviewers (lines 6 and 7 of Table 2). The cost associated with this change was a decrease of 2% in performance [3] , which is considered acceptable, since it also implies including a more significant amount of TBs that represent questions. We suspect that this decrease may be due to the different characteristics of TBs composed of questions, as they end with a higher intonation, instead of a lower intonation, typical of TBs composed of affirmations. Considering that the impact of these three changes was not considered very significant and considering the circumstances that led us to those changes, the model we use for the following experiments includes all of the changes. We refer to it as ProsSegue-ML-MuDi, or simply MuDi.

4.2 Robustness Evaluation

In (Craveiro et al., 2025), despite the usage of a relatively diverse corpus (detailed in (Craveiro and Galdino, 2025)) to train and test their model, since the speech excerpts of the training and test sets are from the same speakers, the model is relatively biased, as speakers' unique voices are present in both sets [4]. Here, we present a robustness evaluation by testing MuDi, the updated model, in different corpora. Table 3 compares the performance of the model in samples from corpora C-ORAL BRASIL I and II (Mello et al., 2012; Mello et al.), NURC-SP Minimum Corpus (Santos et al., 2022), and MuPe-Diversidades' test set (Craveiro and Galdino, 2025). The macro F1 obtained with the MuPe-Diversidades test set is 14% higher than that obtained with NURC-CM, and 4% higher than the one obtained with samples from C-ORAL I and II. Those results show that the model functions well in C-ORAL BRASIL, which is a small and less complex dataset, but functions consistently worse for NURC-CM. The average macro F1 considering all corpora is 69%. Regarding NURC-SP Minimum Corpus, its results could be lower (binary f1-score of 26% and macro f1-score of 61%) due to the poorer quality of the audios and to the presence of occasional overlaps of voices.

4.2.1 ProSegue-ML-MuDi x other methods

Tables 4 and 5 compare the results obtained with MuDi when tested in the same corpora used in other works.

(Craveiro et al., 2024) presents a methodology based on three heuristics that identify pauses and differences of speech rate, which is justified since the lengthening of speech rate at the end of a unit, together with the acceleration at its beginning, is a characteristic of prosodic boundaries among units (Biron et al., 2021). Table 4 shows the comparison between the results they reported and the results we obtained by testing ProsSegue-ML-MuDi in the data samples that they used. The machine-learning-based method shows values that range from a decrease of 6% to an increase of 4% in binary F1. The performance, which was lower than expected, could be explained by the different characteristics

[2]While the interviewees were carefully selected according to gender, age, and region of birth, the interviewers were not controlled in such a manner.

[3]This decrease of 2% may have also been affected by the implementation of a correction on the attribution of labels, which avoided wrongly labeling a sequence of syllables as TBs. This type of error occurred only when the last TB of the reference transcription ended before the ending time of a few syllables, as aligned with UFPAlign. In those cases, before the correction, all of those adjacent syllables were being labeled as "TB" instead of "NB".

[4]Note that this bias will occur for all cases that use samples from MuPe-Diversidades as the test set and samples from MuPe-Diversidades in the training set

Table 3: Table comparing the performance of ProSSegue-ML across different corpora. F1, recall, and precision solely considering TBs are at the left, and the average of TBs and NBs (no boundary) are at the right. Acc. = accuracy, bF1 = binary F1 score, mF1 = macro F1 score.

Test Corpus	Size	Gender	Region	Age range	Education	bF1 / mF1	Acc.	Precision	Recall
Mupe-Diversidades	∼30min	balanced	17 states	20-91	varied	53% / 75%	95%	55% / 76%	51% / 74%
C-oral I and II	∼17min	male	5 states	-	-	44% / 71%	95%	32% / 66%	60% / 83%
NURC-CM	∼17.5 hrs	balanced	SP	25-85	higher	26% / 61%	93%	24% / 60%	30% / 63%
Total Avg	∼18.3hrs	-	-	-	-	41% / 69%	94%	37% / 67%	47% / 73%

Table 4: Overall results of the baseline method applied only to TBs from four inquiries from NURC-MC, compared to the machine learning approach performance applied to the same inquiries.

	SP_EF_156		**SP_DID_242**	
	ProsSegue-Baseline	**ProsSegue-ML**	**ProsSegue-Baseline**	**ProsSegue-ML**
F1	0.18	**0.22**	**0.29**	0.23
p	0.12	**0.17**	**0.22**	0.2
r	**0.38**	0.33	**0.41**	0.29
ser	3.55	**2.28**	02.03	**1.89**
	SP_D2_255		**SP_D2_360**	
	ProsSegue-Baseline	**ProsSegue-ML**	**ProsSegue-Baseline**	**ProsSegue-ML**
F1	0.16	**0.19**	**0.2**	0.17
p	0.11	**0.14**	**0.14**	**0.14**
r	**0.32**	0.27	**0.37**	0.21
ser	3.31	**2.34**	2.92	**2.04**

Table 5: Table comparing the performance of Prossegue-ML vs. models TB-b1 and TB-b2, trained in balanced datasets, published at (Teixeira, 2022) (see Appendix A for composition of the Sample I and Sample II).

Article	Model	Train set	Test set	Accuracy	Specificity	Sensitivity
Teixeira 2022	TB-b1	Sample I	Sample II	87.0%	88.0%	74.0%
Teixeira 2022	TB-b2	Sample II	Sample I	91.0%	92.0%	81.0%
Craveiro et al.2025	ProsSegue-ML	MuPe-Diversidades	Sample II	94.6%	95.2%	65.4%
Craveiro et al.2025	ProsSegue-ML	MuPe-Diversidades	Sample I	95.4%	96.2%	71.8%

of the audios from the Minimum Corpus, suggesting that the model may not be very robust and may benefit from training with more representative samples. It also suggests that the model could be relying on similar features, as we know that the most important feature of ProsSegue-ML-MuDi is the duration of pauses, and that ProsSegue-Baseline uses a rule based on silences to identify boundaries.

(Teixeira, 2022) presents an approach based on an LDA classifier trained with a wider set of features that consider pauses, F0, intensity, speech rate, and rhythm. They present several models, and we selected two models that facilitated the comparison between the works. TB-b1 is a model trained in a balanced sample of Sample I with 8 features related to pauses, F0, and articulation rate. TB-b2 is a model trained in a balanced sample of Sample II with 5 acoustic features, including parameters related to pauses and F0 of units adjacent to terminal breaks. The most relevant feature, with a significantly higher importance than the others, in both models, is the presence of pauses.

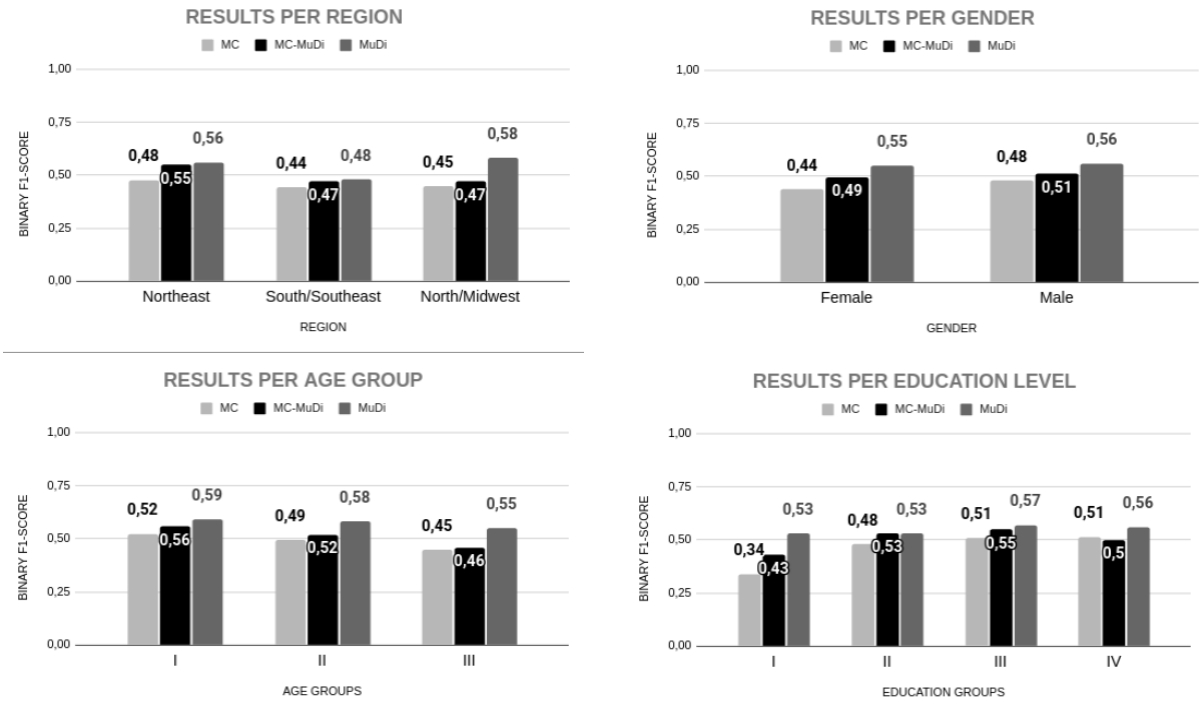
Table 5 shows a comparison of the results that their models obtained and the results obtained with ProSegue-ML-MuDi tested on the same data samples, extracted from corpora C-ORAL BRASIL I and II. ProsSegue-ML-MuDi performs from 4% to 7% higher in accuracy and specificity, while TB-b1 and TB-b2 win by approximately 9% in terms of sensitivity, implying that TB-b1 and TB-b2 are better at not missing TBs, while ProsSegue-ML-MuDi is better at indicating boundaries solely where they actually exist.

4.3 Results per Speaker Profile

Considering how relevant it is that a model is equally effective to all individuals, regardless of their characteristics, we present results per region, gender, age group, and education group, relevant aspects for speech analysis (Craveiro and Galdino, 2025). Figure 1 exhibits results obtained with three different models analyzed under the perspective of each of those aspects. ProsSegue-ML-MC, or simply MC, was trained exclusively with data from the Minimum Corpus, to evaluate whether a much larger dataset (approximately 17 hours) but less diverse (all speakers are from São Paulo and have higher education) suffices for an even performance across diverse speaker profiles. We trained ProsSegue-ML-MuDi with NURC-MC and MuPe-Diversidades to evaluate whether it is worth it to expand a diverse corpus (MuPe-Diversidades) with a larger less diverse corpus (NURC-MC), that is, if the model functions even better for each speaker profile group or if the inequalities of performance also grow.

Regarding gender, ProsSegue-ML-MC performs

Figure 1: Comparison of the performance of the three models (one trained with Minimum Corpus (MC), one trained with MuPe-Diversidades (MuDi), and one trained with a dataset composed of both (MC-MuDi) according to region, gender, age group, and education group, respectively.



4% better for males than for females, and such inequality grew by 3% when we consider results obtained for the model trained exclusively with MuPe-Diversidades train set (MuDi), as reported in (Craveiro et al., 2025). With MC-MuDi, despite the growth of 1% in inequality in respect to MuDi, it is better than MC by 2%. As for region, ProsSegue-ML-MC performed from 8% to 10% worse for speakers from the North and Southeast, when compared to the other regions, which is very curious since its training set contained solely speakers from São Paulo. The difference per region decreased in both new models, reaching 0% to 8% with MC-MuDi, and 1% to 4% with MC, depending on the groups compared. Regarding age group, both new models also favor younger speakers, and the bias grew from 1%-4% with MuDi, to 3%-7% with MC, to 4%-10% with MC-MuDi. It is surprising that the greater bias came from the model trained in both NURC-MC and MuPe-Diversidades. Finally, the difference also grew according to the educational level of the speakers. MC-MuDi and MC seem to strongly favor more educated speakers (groups III and IV), reaching maximum differences of 12% and 17%, respectively, among groups, when speakers with no education are concerned, while MuDi indicated a maximum difference of 4% among different education groups. The performance difference among groups II, III, and IV reaches differences ranging from 0%-5%. Thus, the greater inequality observed, considering all aspects of speaker profiles that we considered, is the bias disfavoring non-educated speakers.

A decrease in performance was already expected in both new models. However, despite the much larger dataset (around 15 hours) and the 19 different speakers comprised in NURC-CM, the difference in performance among different speaker profiles grew, suggesting it may be more valuable to prefer diversity of speakers over quantity of hours of audio available of a more restricted and less diverse set of speakers. Nonetheless, we emphasize that none of those bias evaluations is statistically relevant. The p-value of MC results per region, gender, age group, and education group is, respectively, 0.22, 0.08, 0.19, and 0.34. As for MC-Mudi, the p-values were 0.08, 0.09, 0.09, and 0.61, respectively. We reinforce that a qualitative analysis would be very beneficial to understand what kinds of segmentation errors occurred, as well as to infer limitations of the model. And although we cannot know the reasons for the difference in performance across different speaker profiles or which of their specific aspects were favored or disfavored without

such analysis, we know that certain individual characteristics of the speakers might have influenced the classifier's decisions as the performance of the model varied significantly according to the speaker evaluated. Thus, we definitely need to train using more speakers to improve the model's generalization capability.

5 Final Considerations

In this study, we focused on a low-impact, open-source, automatic prosodic segmentation approach published in (Craveiro et al., 2025). We trained an updated model, MuDi, with the aim of increasing accessibility by updating the version of the forced phonetic aligner, experimenting with the set of features to simplify the requirements of the approach, and assessing the impact of our changes. We explored how MuDi behaves when tested in other corpora, how it compares to other studies, and experimented with the size and diversity of the training set to analyze how bias could be affected. We observed that it maintained a macro F1 above 70% when tested in MuPe-Diversidades teste, and in samples from C-ORAL BRASIL I and II, but had a decrease in performance, reaching 61% of macro F1, when tested in NURC-CM. We also observed that MuDi was probably overspecializing in MuPe-Diversidades, since when we tested MC-MuDi, expanding the dataset with NURC-CM, the performance when testing in MuPe-Diversidades decreased. We could also observe that the results of macro F1 achieved values from 34% to 52% with MC, and values from 43% to 56% with MC-MuDi, with a difference ranging from 1% to 17% in performance, according to the examined speaker profile aspect, values higher than the 1% to maximum 10% obtained training exclusively with MuDi, suggesting that it is worthwhile to prioritize speaker diversity over the quantity of speaking hours. However, it is worth recalling that the bias evaluation did not achieve statistical relevance. For future work, we intend to perform a qualitative evaluation of the segmentation and to test the model on more datasets, including possibly an extended version of MuPe-Diversidades with more speakers from each state and speakers representing all states of Brazil. Also, it would be beneficial to follow the example of the PSST! approach (Roll et al., 2023), a recent study that obtained excellent results, by finetuning ecologically efficient neural models but with large and diverse Brazilian Portuguese datasets that are manually annotated with prosodic segmentation, a work that could begin with NURC-SP Minimum Corpus.

Limitations

We emphasize that our model is still highly misrepresented. It still lacks training with a huge amount of speech excerpts that represent different aspects of Brazilian speech. Our 30 speakers are also a small number to perform statistically relevant diversity tests. This limitation of data is due to the low resource scenario on manual prosodic segmentation annotation, which is a very demanding and long process, limiting us to datasets that were already annotated.

Acknowledgments

This work was carried out at the Center for Artificial Intelligence (C4AI-USP), with support by the São Paulo Research Foundation (FAPESP grant #2019/07665-4) and by the IBM Corporation. This project was also supported by the Ministry of Science, Technology and Innovation, with resources of Law No. 8.248, of October 23, 1991, within the scope of PPI-SOFTEX, coordinated by Softex and published Residence in TIC 13, DOU 01245.010222/2022-44.

References

2025. Error when generating syllphones tier · Issue #19 · falabrasil/ufpalign — github.com. https://github.com/falabrasil/ufpalign/issues/19. Accessed at 02-02-2026.

Sankaranarayanan Ananthakrishnan and Shrikanth S. Narayanan. 2008. Automatic prosodic event detection using acoustic, lexical, and syntactic evidence. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(1):216–228.

Cassio Batista, Ana Larissa Dias, and Nelson Neto. 2022. Free resources for forced phonetic alignment in brazilian portuguese based on kaldi toolkit. *EURASIP Journal on Advances in Signal Processing*, 2022(1):11.

Tirza Biron, Daniel Baum, Dominik Freche, Nadav Matalon, Netanel Ehrmann, Eyal Weinreb, David Biron, and Elisha Moses. 2021. Automatic detection of prosodic boundaries in spontaneous speech. *PLoS ONE*, 16(5):1–21.

Giulia Bossaglia and Lucia de Almeida Ferrari. 2019. The c-oral-brasil project: varied resources for the study of spoken brazilian portuguese. *Journal of speech sciences*.

Meredith Broussard. 2018. *Artificial unintelligence: How computers misunderstand the world*. mit Press.

Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR.

Ken Chen and Mark Hasegawa-Johnson. 2004. How prosody improves word recognition. In *Proc. Speech Prosody 2004*, pages 583–586.

Giovana Craveiro, Caroline Alves, Flaviane Svartman, and Sandra Aluísio. 2025. Machine learning classifiers with acoustic features for prosodic segmentation in brazilian portuguese: A comprehensive evaluation. In *Anais do XVI Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 113–124. SBC.

Giovana Meloni Craveiro and Julio Cesar Galdino. 2025. Diversity in data for speech processing in brazilian portuguese. In *Intelligent Systems*, pages 122–136, Cham. Springer Nature Switzerland.

Giovana Meloni Craveiro, Vinicius Gonçalves Santos, Gabriel Jose Pellisser Dalalana, Flaviane R. Fernandes Svartman, and Sandra Maria Aluísio. 2024. Simple and fast automatic prosodic segmentation of Brazilian Portuguese spontaneous speech. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 1*, pages 32–44, Santiago de Compostela, Galicia/Spain. Association for Computational Lingustics. Available at https://aclanthology.org/2024.propor-1.4/.

Julio Galdino, Sidney Leal, Leticia de Souza, Rodrigo Lima, Antonio Moreira, Arnaldo Candido, Miguel Oliveira, Edresson Casanova, and Sandra Aluísio. 2026a. The impact of prosodic segmentation on speech synthesis of spontaneous speech. In *Intelligent Systems*, pages 547–561, Cham. Springer Nature Switzerland.

Julio Cesar Galdino, Rian Pereira Fernandes, Giovana Meloni Craveiro, Caroline Adriane Alves, Sidney Evaldo Leal, Arnaldo Candido Junior, Flaviane Romani Fernandes-Svartman, and Sandra Maria Aluisio. 2026b. Investigating the effect of automatic prosodic segmentation on speech synthesis for brazilian portuguese. Accepted at Speech Prosody 2026.

Lap Man Hoi, Yuqi Sun, and Sio Kei Im. 2022. An automatic speech segmentation algorithm of portuguese based on spectrogram windowing. In *2022 IEEE World AI IoT Congress (AIIoT)*, pages 290–295.

Jui-Ting Huang, Mark Hasegawa-Johnson, and Chilin Shih. 2008. Unsupervised prosodic break detection in Mandarin speech. In *Proc. Speech Prosody 2008*, pages 165–168.

Je Hun Jeon and Yang Liu. 2009. Semi-supervised learning for automatic prosodic event detection using co-training algorithm. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 540–548, Suntec, Singapore. Association for Computational Linguistics. Available at https://aclanthology.org/P09-1061.

Daniil Kocharov, Tatiana Kachkovskaia, and Pavel Skrelin. 2017. Eliciting Meaningful Units from Speech. In *Proc. Interspeech 2017*, pages 2128–2132.

Cheng-Hsien Lin, Chung-Long You, Chen-Yu Chiang, Yih-Ru Wang, and Sin-Horng Chen. 2019. Hierarchical prosody modeling for Mandarin spontaneous speech. *The Journal of the Acoustical Society of America*, 145(4):2576–2596.

Shimeng Liu, Yoshitaka Nakajima, Lihan Chen, Sophia Arndt, Maki Kakizoe, Mark A. Elliott, and Gerard B. Remijn. 2022. How pause duration influences impressions of english speech: Comparison between native and non-native speakers. *Frontiers in Psychology*, 13.

Heliana Mello, Maryualê Malvessi Mittmann, H. P. Vale, and P.O. Cortes. 2012. Transcrição e segmentação prosódica do corpus C-ORAL-BRASIL: critérios de implementação e validação. In *CORAL-BRASIL I: corpus de referência do português brasileiro falado informal*. Editora UFMG.

Heliana Mello, Tommaso Raso, Lúcia de Almeida Ferrari, and Bruno Neves Rati de Melo Rocha. C-ORAL–Brasil II: Corpus de referência do português brasileiro falado falado formal, mídia e telefone. Available at http://c-oral-brasil.org/c-oral-brasil-ii_N.php. Accessed at 02-02-2026.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Tommaso Raso and Heliana Mello. 2012. The c-oral-brasil i: reference corpus for informal spoken brazilian portuguese. In *International Conference on Computational Processing of the Portuguese Language*, pages 362–367. Springer.

Tommaso Raso, Bárbara Teixeira, and Plínio Barbosa. 2020. Modelling automatic detection of prosodic boundaries for Brazilian Portuguese spontaneous speech. *Journal of Speech Sciences (JOSS)*, 9:105–128.

Nathan Roll, Calbert Graham, and Simon Todd. 2023. PSST! prosodic speech segmentation with transformers. In *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, pages 476–487, Singapore. Association for Computational Linguistics. Available at https://aclanthology.org/2023.conll-1.31/.

Lívia Ruback, Denise Carvalho, and Sandra Avila. 2022. Mitigating bias in machine learning: A socio-technical analysis. *iSys - Brazilian Journal of Information Systems*, 15(1):23:1–23:31.

Vinícius G. Santos, Caroline Adriane Alves, Bruno Baldissera Carlotto, Bruno Angelo Papa Dias, Lucas Rafael Stefanel Gris, Renan de Lima Izaias, Maria Luiza Azevedo de Morais, Paula Marin de Oliveira, Rafael Sicoli, Flaviane Romani Fernandes-Svartman, Marli Quadros Leite, and Sandra Maria Aluísio. 2022. CORAA NURC-SP Minimal Corpus: a manually annotated corpus of Brazilian Portuguese spontaneous speech. In *Proc. IberSPEECH 2022*, pages 161–165.

Bárbara Teixeira, Plínio Barbosa, and Tommaso Raso. 2018. Automatic detection of prosodic boundaries in Brazilian Portuguese spontaneous speech. In *Computational Processing of the Portuguese Language - 13th International Conference, PROPOR 2018*, pages 429–437, Cham. Springer International Publishing.

Bárbara Helohá Falcão Teixeira. 2022. *Detecção automática de fronteiras prosódicas na fala espontânea*. Ph.D. thesis, Universidade Federal de Minas Gerais, Belo Horizonte.

Colin W. Wightman and Mari Ostendorf. 1991. Automatic recognition of prosodic phrases. *[Proceedings] ICASSP 91: 1991 International Conference on Acoustics, Speech, and Signal Processing*, 1:321–324.

A Appendix

A.1 Conflicts in the attribution of labels

In order to test the models in the new corpora, it was necessary to attribute a label to every syllable uttered in each audio. In the process, a few alignment conflicts were found, so we explain how they were dealt with. Ideally, the last syllable of TBs would receive the label "TB"and all others would receive "NB", meaning "no boundary".

However, as we used UFPAlign to identify the initial and final time of the syllables, there are moments when the timestamps of syllables and TBs are not perfectly aligned, that is, UFPAlign might have indicated that a certain syllable ended at 8,09 seconds, for instance as happened in SP D2 360, while the reference annotation indicates that the same syllable was the last syllable of a TB ending in 11,11 seconds, implying that such syllable ended at 11,11 seconds instead of 8,09 seconds. The timestamps of the final times of TBs extracted from the reference annotation are always prioritized, meaning that in cases like this one, this syllable would have received a label "NB", despite being the last syllable of the TB, since it ended before the time indicated at the reference file. And that the syllable that ended near 11,11 seconds, according to UFPAlign's forced alignment, was the one that received the label "TB". In cases where these circumstances were found at the end of the file, the last syllable of the last TB was labeled as "NB". There were also cases where the contrary occurred, that is, UFPAlign's alignment indicated that the final syllable of a TB ended after the final time of the TB as indicated by the annotation. And there were also a few cases where a sequence of syllables had its starting and ending time after the final time of the last TB. In those cases, the syllable that received the label "TB" that corresponds to the indication of the final TB was the syllable with the most approximate time to the end of the TB, and all those "extra" syllables that occurred later were labeled "NB". The code also favors including the first syllable of the following TB at the current TB in case the syllable starts before the current TB ends. Thus, in such cases, the initial syllable of the following TB would be labeled as TB instead of the final syllable of the current TB.

Combining Semantic Embeddings and Knowledge Graphs for Identifying Decision Patterns in Brazilian Judicial Decisions

Gustavo Soares Silva[1], **Omar Andres Carmona Cortes**[1,2], **Fábio Manoel França Lobato**[1,3], **Antonio Fernando Lavareda Jacob Junior**[1],
[1]State University of Maranhão (UEMA), [2]Federal Institute of Maranhão (IFMA),
[3]Federal University of Western Pará (UFOPA),
Correspondence: antoniojunior@professor.uema.br

Abstract

Approaches based solely on textual representations have limitations in capturing structural relations between legal entities, particularly in documents with high lexical similarity. This paper presents ongoing work on a dynamic clustering system for judicial decisions that integrates hybrid representations, combining semantic embeddings from legal-domain Portuguese models with knowledge graphs automatically constructed from documents. The architecture supports incremental clustering and generates cluster justifications using Large Language Models grounded on knowledge graph relations. Preliminary evaluation combines the quantitative metrics Silhouette Score, Davies-Bouldin Index, and Calinski-Harabasz Index.

1 Introduction

Judicial decision-making systems increasingly rely on mechanisms that support consistency across large case volumes, especially in contexts of repetitive litigation(Castro and Mendonça, 2024; Oliveira and Nascimento, 2025). In Brazil, the Judiciary accumulates millions of pending cases (CNJ, 2025), motivating the use of computational methods to assist courts in managing precedents and supporting decision uniformity (Polo et al., 2021). The Brazilian Code of Civil Procedure reinforces this demand by establishing binding precedents (Mentzingen et al., 2024).

Existing approaches frequently rely on textual similarity measures or semantic representations (Silva et al., 2021; Costa et al., 2023). Recent studies have explored hybrid representations that integrate semantic embeddings with Knowledge Graphs (KGs) (Tang et al., 2024; Aguiar et al., 2022), yet questions remain about their performance in dynamic environments and their capacity to support legally meaningful explanations. Methods requiring complete corpus reprocessing face scalability limitations, and the use of Large Language Models (LLMs) to generate justifications raises legal validity concerns (Oliveira and Nascimento, 2025).

This paper presents ongoing work on a dynamic clustering system combining semantic embeddings and automatically constructed knowledge graphs. The system supports incremental clustering and structured justifications, addressing three Research Questions regarding (RQ1) hybrid representations, (RQ2) incremental clustering, and (RQ3) legal validity of generated justifications.

2 Related Work

Text mining in the legal domain presents challenges due to specialized vocabulary, complex syntactic structures, and extensive use of normative references (Polo et al., 2021). For Brazilian Portuguese, BERTimbau (Souza et al., 2020) established itself as a reference model, pre-trained on 2.68 billion tokens. Subsequent domain specialization efforts produced models adapted to legal texts, such as LegalBERT-pt (Silveira et al., 2023) and BumbaBERT (do Carmo, 2024).

In the Brazilian legal context, Silva et al. (2021) compared combinations of textual representation techniques (TF-IDF, Word2Vec) with clustering algorithms (K-Means, Agglomerative, Spectral) on 1,515 initial petitions, finding that TF-IDF with PCA and K-Means produced more coherent clusters. Aguiar et al. (2022) expanded the analysis to 16,000 petitions from a state court, integrating HDBSCAN and BERTimbau with a legal KG. Oliveira and Nascimento (2025) used LLMs to interpret clusters generated from 210,000 labor court documents.

The analysis of related work reveals three limitations. First (L1), all analyzed works process static document sets, without addressing continuous incorporation of new documents. Second (L2),

Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)
Vol. 2, pages 181–185

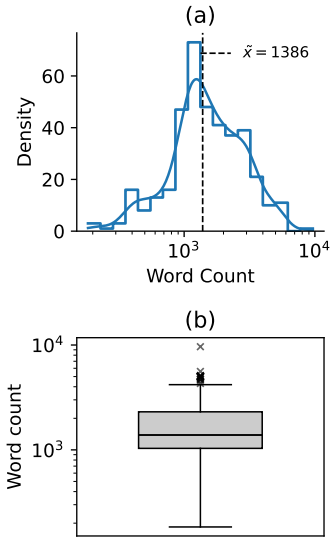


Figure 1: Word count distribution in the corpus: (a) frequency histogram and (b) box plot.

representations capture linguistic patterns but do not explicitly model structural relations such as citations and normative references. Third (L3), when LLMs are used for interpretation, they generate textual descriptions but do not validate the legal consistency of clusters.

This work addresses these gaps by proposing a system that incorporates incremental clustering (L1), hybrid representation combining semantic embeddings and KGs (L2), and LLM-based justification grounded on graph relations (L3).

3 Methodology

This section describes the research methodology following the Data Science Trajectories (DST) framework (Martínez-Plumed et al., 2019), including dataset description, system architecture, and evaluation procedures.

3.1 Business Understanding

The study was conducted in collaboration with the Tribunal de Justiça do Maranhão (TJMA), where decision-pattern identification remains manual and time-consuming. Payroll loan disputes were selected as a pilot domain due to their high volume.

3.2 Data Acquisition and Data Understanding

The corpus comprises 388 judicial decisions from the TJMA (2016–2023) related to payroll loan disputes. Figure 1 shows the word count distribution.

As shown in Figure 1, document length varies substantially, exceeding the token limit of BERT-based models and requiring chunking strategies. Figure 2 compares structural similarity using Jaccard distance over tokens, with semantic similarity using cosine similarity between LegalBERT-pt embeddings. The concentration of document pairs in the upper-right quadrant (structural similarity: 0.86; semantic similarity: 0.94) indicates high lexical and semantic overlap, which may limit the discriminative capacity of clustering approaches based exclusively on textual representations.

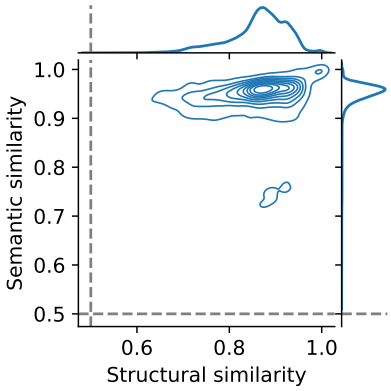


Figure 2: Relationship between structural and semantic similarity across document pairs.

To characterize the corpus's thematic structure, topic modeling was performed using BERTopic with LegalBERT-pt embeddings, UMAP, and HDBSCAN. Figure 3 shows the two-dimensional projection of documents by topic.

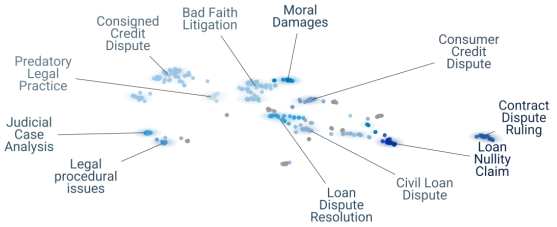


Figure 3: Two-dimensional projection of documents by topics identified via BERTopic.

As shown in Figure 3, the results indicate thematic variation within the corpus, including multiple topics and outliers.

3.3 Data Preparation

Text preprocessing was limited to merging paragraphs fragmented by spurious line breaks and removing excess whitespace; no lowercasing, accent removal, or header stripping was applied, since contextual language models benefit from preserving the original text. Documents were then split into

non-overlapping chunks with a maximum length of 1,500 characters, empirically determined, using recursive character splitting to respect the token limits of extraction models while avoiding duplicate entities in the KG. Chunks and their embeddings are stored in a ChromaDB vector database, which is selected for its local execution capabilities and compliance with data protection requirements. Named entity anonymization is applied after KG extraction to preserve contextual integrity during the extraction stage.

3.4 Modeling

Four embedding models are compared: LegalBERT-pt (Silveira et al., 2023); BERTimbau (Souza et al., 2020); BumbaLM-Embedding, a Qwen3-Embedding-4B variant fine-tuned for the Brazilian legal domain; and a CBOW-based Word2Vec model as a non-contextual baseline. For documents exceeding the token limit of transformer-based models, chunk-level embeddings are averaged to produce a single document vector.

KG construction uses Jurema-7B, a legal-domain LLM for Brazilian Portuguese, hosted on HuggingFace and executed locally, to perform schema-free extraction of entities and relations from each chunk. The resulting graph captures parties, legal provisions, and juridical concepts along with their semantic relations. Graph embeddings are obtained via TransE (Bordes et al., 2013), which models relations as translations in vector space, chosen for its computational efficiency and dimensional compatibility with the textual embeddings.

3.5 Evaluation

Evaluation uses Silhouette, Davies-Bouldin, and Calinski-Harabasz metrics to assess cluster cohesion and separation. The contribution of KG embeddings is assessed by contrasting text-only and hybrid representations across all three metrics. Qualitative evaluation involves legal experts from the TJMA assessing cluster coherence, justification validity, and practical utility through a structured questionnaire.

3.6 Deployment

The system will be delivered as a Docker-containerized microservice that exposes an API that accepts judicial documents and returns cluster assignments with structured justifications, ensuring local execution to ensure data protection compliance.

4 Framework Architecture

The system is organized into five stages. First (a), judicial decisions are segmented into chunks, entities and relations are extracted using an LLM, entity linking resolves coreferences, and the results are incorporated into the KG.

Second (b), semantic embeddings from the domain-specialized model and structural embeddings from TransE are L2-normalized and concatenated. Let $e_t \in \mathbb{R}^n$ denote the textual embedding and $e_g \in \mathbb{R}^m$ the graph embedding. The hybrid representation is defined as

$$e_h = \left[\frac{e_t}{\|e_t\|_2}; \frac{e_g}{\|e_g\|_2} \right] \in \mathbb{R}^{n+m}. \quad (1)$$

Third (c), embeddings are reduced via UMAP with empirically selected target dimensions $d \in \{2, 10, 50\}$, and initial clustering is performed over the reduced vectors using K-Means, HDBSCAN, Agglomerative, and Spectral methods, producing reference clusters $C = \{C_0, \ldots, C_K\}$.

Fourth (d), incremental clustering assigns each new document j to the most similar existing cluster or creates a new one. Let $s_j^{(k)}$ denote the cosine similarity between $e_h^{(j)}$ and the centroid of cluster C_k. Given a threshold θ_s, the document is assigned to C_{k^*} if $\max_k s_j^{(k)} \geq \theta_s$; otherwise, a new cluster C_{K+1} is created. Centroids are updated incrementally without reprocessing the full corpus. The threshold θ_s is initially defined as a user-configurable parameter, allowing legal professionals to control the granularity of the cluster according to operational needs. Data-dependent estimation strategies are planned for future work.

Finally (e), an LLM generates cluster-level justifications grounded in the subgraph of entities and relations shared by the cluster members.

5 Preliminary Results

This section presents preliminary results from experiments with text-only representations, which serve as a baseline for subsequent evaluation of hybrid representations. To obtain a single ranking that balances all three metrics, each configuration is ranked independently by Silhouette Score, Davis-Bouldin Index, and Calinski-Harabasz Index, and the mean of the three positional ranks is computed.

Embedding	Reducer	Algorithm	k	S ↑	DB ↓	CH ↑	Mean Rank ↓
BumbaLM	UMAP(d=2)	HDBSCAN	5	0.752	0.289	8,232.49	37.7
BERTimbau	UMAP(d=2)	HDBSCAN	5	0.745	0.309	3,540.24	94.7
BumbaLM	UMAP(d=10)	HDBSCAN	5	0.723	0.373	3,743.54	123.0
CBOW	UMAP(d=10)	HDBSCAN	5	0.719	0.378	3,212.93	146.3
BumbaLM	UMAP(d=2)	HDBSCAN	10	0.663	0.382	3,984.78	148.3

Table 1: Top-5 clustering configurations by average internal positional rank across Silhouette Score (S), Davies-Bouldin Index (DB), and Calinski-Harabasz Index (CH)

Table 1 presents the top-5 configurations by this aggregated criterion.

The configuration that achieved the highest overall ranking according to the aggregated metric combines BumbaLM embeddings with UMAP ($d = 2$) for dimensionality reduction and HDBSCAN ($k = 5$) for clustering, yielding S = 0.752, DB = 0.289, and CH = 8232.49. BumbaLM appears in three of the five top-ranked configurations, which may indicate that the legal-domain fine-tuning of this embedding model contributes to more discriminative representations in this context.

The presence of CBOW among the top configurations indicates that simpler embedding approaches can still produce competitive results when combined with suitable dimensionality reduction and clustering techniques. Additionally, all top-performing configurations rely on the combination of UMAP and HDBSCAN.

These results indicate that evaluation metrics do not always favor the same configurations, reinforcing the need for qualitative expert assessment and KG-based representations.

6 Preliminary Conclusions

This paper presented ongoing work on a dynamic clustering system that integrates semantic embeddings and KGs, addressing limitations related to static document processing (L1), text-only representations without structural relations (L2), and cluster interpretation without legal validation (L3).

Preliminary results with text-only representations indicate that legal-domain fine-tuned models, particularly BumbaLM, produce more discriminative embeddings for clustering judicial decisions, and that density-based clustering (HDBSCAN) with UMAP consistently outperforms partition-based alternatives. However, disagreements among internal validation metrics highlight that quantitative evaluation alone is insufficient to assess legally meaningful cluster quality, motivating both the integration of KG-based structural features and qualitative expert validation.

Future stages involve implementing and evaluating hybrid representations, incremental clustering, and extending to additional legal domains.

Limitations

This work has limitations inherent to its current stage. Although the proposed pipeline is designed to be domain-agnostic and applicable across legal jurisdictions, empirical validation is currently limited to 388 documents from a single court addressing payroll loan disputes. The high semantic overlap in this corpus makes it a challenging testbed where KG-based differentiation is most needed, but evaluation across additional domains is required to confirm the approach's generalizability. An expanded corpus has been requested from the court.

Also, a few components remain pending: the empirical comparison between text-only and hybrid representations, the assessment of incremental clustering sensitivity to θ_s, and qualitative validation by legal experts.

The schema-free KG extraction relies on Jurema-7B without independent validation of entity and relation quality, and the entity linking step has not been formally evaluated. Similarly, the LLM-produced justifications have not yet been assessed for legal adequacy; while grounding on KG relations is designed to mitigate hallucination, its effectiveness remains an open question.

Acknowledgments

This study was supported by the National Council for Scientific and Technological Development (CNPq) - DT-303031/2023-9; by the Maranhão Foundation for Research and Scientific and Technological Development; the Financing Agency for Studies and Projects (FINEP) – (ProAmazonia - 2373/24 - CTCCA-II); and by Technical Cooperation Agreement Nº. 02/2021 (case Nº. 38328/2020-TJ/MA).

References

André Aguiar, Raquel Silveira, Vasco Furtado, Vládia Pinheiro, and João A. Monteiro Neto. 2022. *Using Topic Modeling in Classification of Brazilian Lawsuits*, page 233–242. Springer International Publishing.

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Durán, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. In *Proceedings of the 27th International Conference on Neural Information Processing Systems*, volume 2 of *NIPS'13*, page 2787–2795. Curran Associates Inc.

Marcella Queiroz de Castro and Ana Régia Mendonça. 2024. PLN e segurança jurídica identificação de divergências jurisprudenciais com processamento de linguagem natural. In *Proceedings of the 15th Brazilian Symposium in Information and Human Language Technology*, pages 457–462, Belém do Pará, Brazil. Association for Computational Linguistics.

CNJ. 2025. Justiça em números.

José Alfredo F. Costa, Nielsen Castelo D. Dantas, and Esdras Daniel S. A. Silva. 2023. Evaluating text classification in the legal domain using bert embeddings. In *Intelligent Data Engineering and Automated Learning – IDEAL 2023*, pages 51–63, Cham. Springer Nature Switzerland.

Fabrício Almeida do Carmo. 2024. Representações *Embeddings* Orientadas à Linguagem Jurídica Brasileira. Mestrado em engenharia da compuatação e sistemas, Universidade Estadual do Maranhão, São Luís - MA.

Fernando Martínez-Plumed, Lidia Contreras-Ochando, Cesar Ferri, José Hernández-Orallo, Meelis Kull, Nicolas Lachiche, María José Ramírez-Quintana, and Peter Flach. 2019. Crisp-dm twenty years later: From data mining processes to data science trajectories. *IEEE Transactions on Knowledge and Data Engineering*, 33(8):3048–3061.

Hugo Mentzingen, Nuno António, Fernando Bacao, and Marcio Cunha. 2024. Textual similarity for legal precedents discovery: Assessing the performance of machine learning techniques in an administrative court. *International Journal of Information Management Data Insights*, 4:100247–100247.

Raphael Souza de Oliveira and Erick Giovani Sperandio Nascimento. 2025. Analysing similarities between legal court documents using natural language processing approaches based on transformers. *PLOS ONE*, 20(4):e0320244.

Felipe Polo, Gabriel Mendonça, Kauê Parreira, Lucka Gianvechio, Peterson Cordeiro, Jonathan Ferreira, Leticia Lima, Antônio Maia, and Renato Vicente. 2021. Legalnlp - natural language processing methods for the brazilian legal language. In *Anais do XVIII Encontro Nacional de Inteligência Artificial e Computacional*, pages 763–774, Porto Alegre, RS, Brasil. SBC.

Ingrid L. A. da Silva, Rafael Ferreira Mello, Péricles B. C. Miranda, André C. A. Nascimento, Isabel W. S. Maldonado, and José L. M. Coelho Filho. 2021. Assessment of text clustering approaches for legal documents. In *Anais do XVIII Encontro Nacional de Inteligência Artificial e Computacional (ENIAC 2021)*, ENIAC 2021, page 37–48. Sociedade Brasileira de Computação.

Raquel Silveira, Caio Ponte, Vitor Almeida, Vládia Pinheiro, and Vasco Furtado. 2023. Legalbert-pt: A pretrained language model for the brazilian portuguese legal domain. In *Intelligent Systems*, pages 268–282, Cham. Springer Nature Switzerland.

Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. 2020. Bertimbau: Pretrained bert models for brazilian portuguese. In *Intelligent Systems*, pages 403–417, Cham. Springer International Publishing.

Yanran Tang, Ruihong Qiu, Yilun Liu, Xue Li, and Zi Huang. 2024. Casegnn: Graph neural networks for legal case retrieval with text-attributed graphs. In *Advances in Information Retrieval*, pages 80–95, Cham. Springer Nature Switzerland.

Development and Evaluation of a Hybrid Information Retrieval System Applied to the Brazilian Legal Domain

Ana Carolina C. Bessa[1], Fábio M. F. Lobato[1,2], Antonio F. L. J. Junior[1]

[1]State University of Maranhão, Maranhão, Brazil,
[2]Federal University of Western Pará, Pará, Brazil

Correspondence: anabessa@aluno.uema.br,, fabio.lobato@ufopa.edu.br, antoniojunior@professor.uema.br

Abstract

The need for tools to manage processes, automate tasks, and speed up the judicial system justifies enhancing traditional Information Retrieval systems, which are often hindered by vocabulary mismatches and lengthy legal texts. Although Transformer-based models capture semantic nuances, they face input size limitations, making it challenging to process long texts without information loss. In this work, we introduce and evaluate a hybrid system for the legal domain that combines the BM25L algorithm with the BumbaLM language model. The system was evaluated using legal judgment summaries from TJMA. The experiments revealed that the standalone semantic model outperformed the hybrid approach. The lexical component struggled with natural-language and conceptual queries, resulting in false positives that degraded the hybrid system's overall performance.

1 Introduction

The volume of documents produced by judicial institutions continues to grow. According to the Brazilian National Council of Justice (CNJ), in its report *Justice in Numbers*[1], the year 2025 ended with 74,756,005 pending cases. This makes it difficult to organize and access information. With the current technological transformation from Industry 4.0, the legal sector has invested in intelligent tools to automate repetitive tasks and reduce procedural delays (Nascimento, 2024). In this sense, Information Retrieval (IR) has gained attention. IR can be defined as the process of finding relevant documents from large amounts of unstructured data. It can be used to locate documents that satisfy a user's information needs (Vitório et al., 2025). In the legal field, this is difficult because of the nature of the documents. They often contain technical language, lengthy texts, and mismatches in vocabulary. Terms in the user's query may not exactly match those in relevant documents (Moreira, 2024).

One technique to overcome this challenge is to use Transformer-based neural models, such as Sentence-BERT (SBERT). These models process entire text sequences at once to capture word relationships. However, using these architectures alone in the legal domain has limitations. Input size restrictions usually limit them to 512 tokens, where a token is roughly a word or character sequence the model can process at once. This requires truncating long texts, possibly losing relevant information in lengthy proceedings. Probabilistic algorithms like Best Matching 25 (BM25), which score documents by matching query terms, perform well in document retrieval and identifying technical terminology. Semantic models may underperform here due to noise or overgeneralization. Therefore, integrating these methods in hybrid systems can combine the semantic power of language models with the efficiency of lexical correspondence.

This research proposes the development of a hybrid legal information retrieval system. It combines the BM25 algorithm with the BumbaLM language model. The choice of BM25L is justified by its ability to handle lengthy documents. This matches the complex structure of legal proceedings. BumbaLM, a model trained on Portuguese legal data, serves as the semantic component. It addresses vocabulary differences and captures conceptual relationships.

2 Related Works

The legal domain presents unique challenges due to its specialized terminology and complex documents. Prior work by Vitório et al. (2025) compared 12 SBERT models to traditional Okapi BM25 and BM25L baselines in the Brazilian legislative context. Neural models improve semantic understanding. Selection of the appropriate BM25 variant is essential for baseline performance. In con-

[1]https://www.cnj.jus.br/wp-content/uploads/2025/11/justica-em-numeros-2025.pdf

Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)
Vol. 2, pages 186–190
April 13-16, 2026

trast, this research focuses on jurisprudential data, summaries of court rulings with different structures. It aims to find the best hybrid approach between BM25 and BERT-based models.

Kodri et al. (2025) introduce the Fine-Hybrid system, which combines BM25 with an SBERT model adapted to a tax corpus. Their results show that domain adaptation improves the model's ability to capture legal nuances. While Kodri et al. (2025) focuses on the tax domain, this study uses a 4-billion-parameter embedding model. It extends hybridization to general jurisprudence.

Another relevant work for the context is Fernandes et al. (2025). The authors introduced the JutisTCU dataset, with more than 16,000 documents from the Brazilian Federal Court of Accounts (TCU). Their experiments showed that integrating semantic methods based on OpenAI and BERT models improves case-law retrieval. Our work differs in that it uses an open-weight model, BumbaLM, which allows courts to run the system locally.

Baban Gain et al. (2019) and Kim et al. (2022b) describe the use of BM25 combined with BERT in Competition on Legal Information Extraction and Entailment (COLIEE) tasks for information retrieval in legal documents in other languages. The hybrid system developed focuses on adapting these architectures to the particularities of Brazilian Portuguese.

In summary, related work shows that combining BERT-based models with algorithms such as BM25 achieves better results for legal IR tasks than using either method alone.

3 Methodology

This section describes the experiments conducted to verify the efficiency of the hybrid model relative to individual information retrieval techniques, as well as the database and evaluation metrics used.

3.1 Data collection

The hybrid system experiments were conducted using a set of legal judgments from the Maranhão Court of Justice (TJMA). The dataset includes 100,000 records of final decisions on various cases, with details such as identification numbers (ID), unique case numbers, district, chamber, CNJ classification (ID and name), summary, and content. For the experiments, only the ID, a unique small number for each document, and the judgment summaries, written by the judges, were selected because their smaller size fit within the 512-token limit of the chosen language model.

The corpus underwent preprocessing to standardize words to lowercase, remove invalid symbols and special characters (e.g., "\n"), and filter out stopwords. This reduced text noise, leaving only the information necessary to improve the system's information retrieval performance.

3.2 BM25 Algorithm

The BM25L algorithm was chosen for exact-term-matching retrieval. This choice suits the legal area where document lengths vary greatly. BM25 is a widely used lexical search algorithm, known for its high efficiency and for requiring fewer computational resources than pre-trained large language models.

BM25L improves upon BM25, which overly penalizes long documents due to term-frequency saturation. BM25L adjusts for length normalization, so document relevance isn't underestimated by length (Kim et al., 2022a). Rare terms retain their discriminatory weight, even in lengthy legal documents.

3.3 BumbaLM

The vector representation was performed using the BumbaLM embedding language model (Carmo et al., 2023). BumbaLM was selected because it was trained on a collection of Portuguese legal documents from the Court of Justice, and it best matched the characteristics of the current dataset. It is important to note that BumbaLM is distributed solely as an open-weight model, enabling courts and legal institutions to run the system locally and ensuring data security during inference. However, the model's full training dataset and source code cannot be made open source due to strict privacy policies and confidentiality restrictions associated with the sensitive, real-world judicial documents used to train it.

Given the characteristics of the chosen model, the texts were truncated to a maximum of 512 tokens due to the model's context window. As a result, priority was given to using sentence summaries prepared by the judges, as they are shorter compared to the full text. The embeddings were created using the Mean Pooling method (averaging the last hidden layer), followed by L2 (Euclidean) normalization to ensure all vectors had unit length. These vectors were then stored in ChromaDB to eliminate the need to generate embeddings in sub-

sequent runs, with cosine distance defined as the chosen similarity metric.

3.4 Proposed system

The hybrid system combines lexical and semantic components for the IR task. The challenge at this stage was the incompatibility of the scoring scales: BM25L produces unlimited scores $[0, \infty]$, based on term frequency, while cosine similarity operates in the range $[-1, 1]$.

To address this issue, the approach used parallel retrieval, in which for each query q, the Top-K documents from each component are retrieved separately. The Min-Max Scaling technique was applied to the raw scores from each candidate list, normalizing them to the range [0, 1]. The final score S_f for each document d was computed using Equation 1.

$$S_f(d, q) = \alpha \cdot S'_1(d, q) + (1 - \alpha) \cdot S'_2(d, q) \quad (1)$$

where S' represents the normalized score, with S'_1 corresponding to BM25L and S'_2 for BumbaLM, and α is the control hyperparameter.

In the comparative experiments, α varied between 0.0 (purely semantic), 1.0 (purely lexical), and intermediate values (hybrid), with the value $\alpha = 0.5$ adopted as the baseline for evaluating the balance between the techniques.

3.5 Evaluation metrics

To assess the hybrid system and the individual techniques, the metrics used were Mean Average Precision (MAP), Mean Reciprocal Rank (MRR), and Mean R-Precision (MRP).

MAP is the main metric for overall quality because it considers the ranking order of all relevant documents retrieved and penalizes the system if an important document appears in lower positions (Zhang and Zhang, 2009). It is calculated by averaging the Average Precision (AP) scores for the set of queries Q. The AP of a query is given by Equation 2.

$$\text{MAP} = \frac{\sum_{k=1}^{n} (P(k) \times rel(k))}{\text{number of relevant documents}} \quad (2)$$

where $P(k)$ represents the precision at cutoff k, and $rel(k)$ is a binary function (1 if the document at position k is relevant, 0 otherwise.

MRR measures the system's ability to find the correct answer as quickly as possible (Caseli and Nunes, 2024). This metric assesses the position of the first relevant document in the list, as shown in Equation 3.

$$\text{MRR} = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{rank_i} \quad (3)$$

where $rank_i$ indicates the position of the first relevant result for the query i.

The MRP metric measures precision at position R, where R is the total number of relevant documents for the current query (Beitzel et al., 2009), as in Equation 4.

$$\text{R-Precision} = \frac{\text{Relevant documents}}{R} \quad (4)$$

The experimental results considering the materials and methods presented are given in the following Section.

4 Results and Discussion

The experiment used 10 queries to retrieve specific results from the database, each linked to the document IDs it referenced. The queries were manually created by a person after analyzing various documents in the database, with each query serving as the main topic of the case, for example: "Property damage caused by a power grid failure". Manual query creation was chosen in this study to ensure that the queries were contextually meaningful and directly relevant to real-world scenarios observed in the dataset. Although this method generated precise, targeted queries, it limited the number of queries because of the time and expertise required for selection. It is also important to note that the dataset does not include annotations indicating the relevance of each judgment to its query at the time of this study. Table 1 shows the results from the models alone and with hybrid techniques.

	Metrics		
Model	MAP	MRP	MRR
BM25L	0,144	0,144	0,1
BumbaLM	**0,583**	**0,583**	**0,3**
Hybrid System	0,335	0,335	0,1

Table 1: Performance comparison between recovery strategies.

The results showed that using only BumbaLM was better than the other scenarios. The semantic model achieved a MAP and MRR of 0.583, meaning that, on average, the first relevant document was ranked between 1st and 2nd place. In comparison, BM25L had the lowest performance (0.144), likely

due to difficulty handling the queries' textual structure. The test queries, formulated in natural language and using conceptual terms, did not produce exact matches in the indexed summaries. Since BM25L relies on exact term frequency, it failed to retrieve documents that used synonyms or paraphrases. Conversely, BumbaLM showed strong generalization. Even without exact word matches, the generated embeddings effectively captured the search intent.

The queries used in the test, formulated in natural language and using conceptual terms, did not produce exact matches in the indexed summaries. BM25L relies on the frequency of exact terms, so it failed to retrieve documents that used synonyms or paraphrases. Even without exact word matches, the generated embeddings captured the search intent.

The hybrid system performed worse than the isolated semantic model, averaging the performance of BumbaLM and BM25L. This combination caused interference in the results by giving too much weight to a low-performing component. Using an average fusion coefficient (0.5), the system allowed BM25L false positives (irrelevant documents with high lexical scores) to overshadow the relevant documents identified by BumbaLM.

One point observed is the exact match between MAP and MRR values across all configurations. This can be explained because, for the set of queries tested, the retrieval functioned like searching for a known item, where the position of the first, and possibly only, relevant document in the template determined the average precision metric.

Therefore, the results showed that, for the evaluated corpus and queries, the BumbaLM language model performed best, and hybridization was harmful in this case. One solution to this issue is to work with a properly annotated dataset, including relevance notes for the judgments, and to formulate queries by legal professionals or LLMs, as presented in the works of Fernandes et al. (2025) and Vitório et al. (2025). Another option is to decrease the weight of the lexical component, using BM25L solely as a tiebreaker.

5 Conclusion

In this study, we presented a hybrid IR system that combines the BM25L algorithm and the BumbaLM model to assess its effectiveness for IR in legal documents. Compared to other work in this area, this work stands out for using an open-weight model with a large number of parameters. The results showed that the hybrid system performed worse than using only BumbaLM for the retrieval task. This is likely due to the structure of the search queries and the absence of relevance indicators in the judgments. Providing a set of annotated data and queries created by lawyers could improve the performance of the hybrid system compared to using each strategy alone.

The experimental results provide some clues into the performance of information retrieval using purely semantic models, purely lexical models, and a hybrid approach, and may guide broader experiments aimed at informing the development of efficient IR systems that meet the intrinsic needs of the legal domain.

For future work, the experiment will be conducted using an annotated database by legal experts and will include the development of an intuitive, accessible Graphical User Interface (GUI) that allows legal professionals to formulate queries in natural language and view the retrieved sentences, with highlights of the passages that contributed to lexical or semantic relevance. Usability tests and qualitative evaluations will be conducted with end users. The goal is to collect human feedback to verify whether the system retrieves the correct documents and is perceived as useful in real-world operational scenarios.

Limitations

As mentioned in the section 4, the limitation of this study is the lack of datasets with scores and answers for each document. In the experiments conducted, it was not possible to achieve the system's correct performance because the most relevant document to the query was not identified. Additionally, the queries were not formulated by legal experts, making it difficult to understand how searches are performed.

Acknowledgments

This study was supported by the National Council for Scientific and Technological Development (CNPq) - DT-303031/2023-9; by the Maranhão Foundation for Research and Scientific and Technological Development; the Financing Agency for Studies and Projects (FINEP) – (ProAmazonia - 2373/24 - CTCCA-II); and by Technical Cooperation Agreement Nº. 02/2021 (case Nº. 38328/2020-TJ/MA).

References

Baban Gain, Dibyanayan Bandyopadhyay, Tanik Saikh, and Asif Ekbal. 2019. Iitp in coliee@icail 2019: Legal information retrieval using bm25 and bert.

Steven M. Beitzel, Eric C. Jensen, and Ophir Frieder. 2009. *Average R-Precision*, pages 195–195. Springer US, Boston, MA.

Fabrício Carmo, Ferdinando Serejo, Antonio Jacob Junior, Ewaldo Santana, and Fábio Lobato. 2023. Embeddings jurídico: Representações orientadas à linguagem jurídica brasileira. In *Anais do XI Workshop de Computação Aplicada em Governo Eletrônico*, pages 188–199, Porto Alegre, RS, Brasil. SBC.

H. M. Caseli and M. G. V. Nunes, editors. 2024. *Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português*, 3 edition. BPLN.

Leandro Carísio Fernandes, Leandro dos Santos Ribeiro, Marcos Vinícius Borela de Castro, Leonardo Augusto da Silva Pacheco, and Edans Flávius de Oliveira Sandes. 2025. Juristcu: A brazilian portuguese information retrieval dataset with query relevance judgments. *arXiv preprint arXiv:2503.08379*.

Gyeongmin Kim, Minseok Kim, and Jaechoon Jo. 2022a. Enhancing code similarity with augmented data filtering and ensemble strategies. *JOIV: International Journal on Informatics Visualization*, 6(3):676–680.

Mi-Young Kim, Juliano Rabelo, Kingsley Okeke, and Randy Goebel. 2022b. Legal information retrieval and entailment based on bm25, transformer and semantic thesaurus methods. *The Review of Socionetwork Strategies*, 16(1):157–174.

Wan Ahmad Gazali Kodri, Muhammad Haris, and Rifqi Fitriadi. 2025. Fine-hybrid: Integration of bm25 and finetuned sbert to enhance search relevance. *Teknika*, 14(2):213–222.

Viviane P. Moreira. 2024. Recuperação de informação. In H. M. Caseli and M. G. V. Nunes, editors, *Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português*, 3 edition, book chapter 21. BPLN.

Iury Gregory Chaves do Nascimento. 2024. A inteligência artificial no sistema judiciário trabalhista brasileiro.

Douglas Vitório, Ellen Souza, José Antônio dos Santos, André Carlos Ponce de Leon Ferreira, Adriano LI Oliveira, Nádia FF da Silva, and 1 others. 2025. Bm25 x vila sésamo: avaliando modelos sentence-bert para recuperaçao de informaçao no cenário legislativo brasileiro. *Linguamática*, 17(1):17–33.

Ethan Zhang and Yi Zhang. 2009. *Average Precision*, pages 192–193. Springer US, Boston, MA.

Viés e Justiça em Modelos de Linguagem: Evidências de Uma Literatura Linguística, Social e Culturalmente Assimétrica

Vitória P. Firmino[1], Bruno M. Nogueira[1], Valéria Q. dos Reis[1]
[1]Universidade Federal de Mato Grosso do Sul,
Correspondence: {vitoria.firmino,bruno.nogueira,valeria.reis}@ufms.br

Resumo

O uso crescente de Grandes Modelos de Linguagem (LLM) tem ampliado preocupações relacionadas a viés social e justiça algorítmica. Este trabalho apresenta uma Revisão Sistemática da Literatura de 60 estudos publicados entre 2020 e 2025, analisando estratégias de mitigação, métricas de avaliação, tipos de discriminação e idiomas considerados. Os resultados indicam forte predominância de avaliações em língua inglesa, foco desproporcional no viés de gênero tratado de forma binária e maior ênfase em diagnóstico do que em mitigação. Observa-se ainda escassez de análises interseccionais, multilíngues e orientadas a cenários reais de uso, evidenciando lacunas metodológicas e socioculturais na literatura atual.

1 Introdução

Os Grandes Modelos de Linguagem (*Large Language Models* – LLMs) têm transformado profundamente as tecnologias de linguagem natural e sido amplamente incorporados a sistemas que apoiam decisões em domínios sensíveis, como educação, recrutamento e comunicação digital (Brown et al., 2020). Contudo, esse avanço é acompanhado por riscos sociais relevantes, em especial a capacidade desses modelos de aprender, reproduzir e amplificar vieses sociais presentes nos dados de treinamento, majoritariamente oriundos da internet (Bender et al., 2021; Sheng et al., 2021).

A investigação empírica de vieses em modelos de linguagem antecede os LLM. Trabalhos seminais demonstraram que representações distribuídas capturam associações estereotipadas semelhantes às humanas (Caliskan et al., 2017) e que tais associações podem ser amplificadas pelos modelos, motivando técnicas iniciais de mitigação em embeddings (Bolukbasi et al., 2016). Esses estudos consolidaram o viés como um fenômeno estrutural em sistemas de linguagem.

Essa preocupação é reconhecida explicitamente no artigo do GPT-3 (Brown et al., 2020), que discute como vieses nos dados de treinamento podem levar à geração de conteúdo discriminatório. Abordagens sociotécnicas ampliam essa análise ao argumentar que o viés algorítmico emerge não apenas dos dados, mas também de escolhas de projeto, objetivos de otimização e padrões de uso, evidenciando limitações de noções estritas de igualdade e a necessidade de perspectivas orientadas à equidade (Mehrabi et al., 2021). A ausência de consenso sobre definições formais de justiça algorítmica reforça essa complexidade, uma vez que diferentes critérios técnicos podem ser incompatíveis entre si e percebidos de forma distinta socialmente (Saxena et al., 2018).

Estudos recentes em Processamento de Linguagem Natural indicam ainda que o viés varia entre idiomas e contextos culturais, com modelos tendendo a favorecer grupos dominantes em cada contexto linguístico, o que limita a generalização de métricas e estratégias concebidas como universais (Levy et al., 2023). Essa constatação dialoga com a literatura das Ciências Sociais, que compreende a discriminação como um fenômeno historicamente situado e culturalmente mediado (Bourdieu, 1991; Tilly, 1998; Lamont et al., 2016).

Diante desse cenário, este trabalho conduz uma Revisão Sistemática da Literatura (RSL) com o objetivo de identificar, classificar e analisar criticamente abordagens para avaliação e mitigação de vieses sociais em modelos de linguagem, adotando uma perspectiva sensível ao idioma, ao contexto sociocultural e ao tipo de discriminação analisado.

2 Trabalhos Relacionados

Revisões amplas sobre justiça algorítmica em aprendizado de máquina são apresentadas por Mehrabi et al. (2021) e Tang et al. (2023), que sistematizam fontes de viés, definições formais de equidade,

Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)
Vol. 2, pages 191–200

métricas e estratégias de mitigação em diferentes domínios. Esses trabalhos oferecem bases conceituais e normativas robustas, articulando perspectivas técnicas e filosóficas, mas mantêm um escopo geral para aprendizado de máquina, anterior ou pouco específico à consolidação dos grandes modelos de linguagem, sem tratar o idioma como dimensão analítica central.

Em contextos aplicados, estudos como o de Fabris et al. (2025) analisam a discriminação algorítmica em processos de recrutamento, organizando métricas e estratégias de mitigação ao longo do ciclo de desenvolvimento e evidenciando limitações recorrentes da literatura, como o predomínio de datasets em língua inglesa, o foco em gênero binário e a escassez de dados sobre outros grupos protegidos. Embora restrito a um domínio específico, esse trabalho reforça a natureza sociotécnica do viés em sistemas baseados em linguagem.

Mais recentemente, Gallegos et al. (2024) apresentam uma síntese abrangente sobre viés e justiça em LLM, distinguindo danos representacionais e alocacionais e propondo taxonomias para métricas, datasets e estratégias de mitigação. No entanto, trata-se de uma revisão narrativa, sem protocolo sistemático explícito, que também não analisa de forma estruturada em quais idiomas as avaliações são conduzidas nem como diferentes tipos de discriminação social são abordados em cada contexto linguístico.

Em conjunto, esses trabalhos demonstram avanços significativos na conceituação e avaliação da justiça algorítmica, mas evidenciam a ausência de revisões sistemáticas que integrem rigor metodológico, avaliação de qualidade e sensibilidade sociolinguística. Diante dessas lacunas, o presente trabalho conduz uma Revisão Sistemática da Literatura com protocolo explícito e replicável, focada em modelos de linguagem, analisando métricas de justiça e estratégias de mitigação à luz dos idiomas avaliados, dos tipos de discriminação social considerados e das limitações metodológicas da literatura recente.

3 Metodologia

A condução desta Revisão Sistemática da Literatura (RSL) seguiu as diretrizes propostas por Kitchenham (2007) e suas atualizações metodológicas (Carrera-Rivera et al., 2022), que enfatizam rigor, transparência e replicabilidade. Essas diretrizes orientaram a definição do protocolo de pesquisa, da estratégia de busca, dos critérios de seleção e da avaliação da qualidade dos estudos. A revisão foi estruturada em torno de cinco perguntas de pesquisa, investigando: (RQ1) as estratégias de mitigação de viés aplicadas a modelos de linguagem; (RQ2) os métodos e métricas utilizados para avaliação de justiça; (RQ3) os tipos de discriminação social abordados; (RQ4) os idiomas nos quais essas abordagens têm sido avaliadas; e (RQ5) as principais tendências, limitações e lacunas identificadas na literatura recente. Essas questões foram formuladas para capturar não apenas aspectos técnicos da avaliação e mitigação de viés, mas também dimensões linguísticas e sociais frequentemente negligenciadas na literatura.

3.1 Critérios PICOC

O escopo da revisão foi definido com base no modelo *PICOC*, considerando como população os modelos de linguagem; como intervenção estratégias, técnicas ou análises voltadas à mitigação, mensuração ou diagnóstico de viés algorítmico; como comparação outras abordagens, referências ou a ausência de intervenção; como resultado métricas de justiça, análises de viés e evidências empíricas, incluindo a avaliação da eficácia de estratégias de mitigação; e como contexto pesquisas científicas nas áreas de Ciência da Computação, Engenharia de Software e Processamento de Linguagem Natural.

3.2 Bases de Dados e Estratégia de Busca

Os estudos primários foram recuperados exclusivamente da base **Scopus**, selecionada por sua ampla cobertura de periódicos e conferências relevantes nas áreas de Computação e Engenharia. A string de busca foi construída a partir dos elementos do modelo PICOC e aplicada aos campos de título, resumo e palavras-chave:

> TITLE-ABS-KEY ((“language models” OR “large language models” OR “LLM”) AND (“algorithm* bias*” OR “discrimination*” OR “fair*” OR “unfair*” OR “algorithmic fairness”) AND (“experiment*” OR “empirical evaluation” OR “case study” OR “implementation” OR “benchmarking”)) AND PUBYEAR > 2019 AND PUBYEAR < 2026

O recorte temporal entre 2020 e 2025 foi adotado para capturar estudos contemporâneos à consolidação dos grandes modelos de linguagem. Os registros foram exportados no formato *BibTeX* em 26 de maio de 2025.

A gestão da revisão — incluindo seleção, aplicação de critérios e extração de dados — foi realizada com o apoio da plataforma **Parsifal**[1], assegurando rastreabilidade e consistência metodológica.

3.3 Critérios de Inclusão e Exclusão

Foram incluídos estudos primários revisados por pares que investigam vieses sociais ou justiça algorítmica em modelos de linguagem, apresentem evidências empíricas e avaliem métricas de justiça ou estratégias de mitigação relacionadas a atributos protegidos, publicados entre 2020 e 2025.

Foram excluídos estudos secundários, trabalhos com foco exclusivamente técnico (como eficiência, arquitetura ou escalabilidade), pesquisas que abordam ética ou governança de forma genérica sem foco em viés social, bem como estudos cuja metodologia não fosse adequada para responder às perguntas de pesquisa definidas.

3.4 Processo de Seleção dos Estudos

O processo de seleção ocorreu em três etapas: (i) triagem por título e resumo; (ii) leitura completa dos estudos pré-selecionados; e (iii) aplicação de *snowballing* a partir das referências dos estudos incluídos.

3.5 Avaliação da Qualidade dos Estudos (QA)

A qualidade metodológica dos estudos primários foi avaliada conforme as recomendações de Kitchenham (2007), com o objetivo de mitigar vieses de seleção e apoiar a interpretação crítica dos resultados. A avaliação foi realizada por meio de uma listagem estruturada baseada em quatro critérios amplamente utilizados em Engenharia de Software e Computação empírica: **Relato**, **Rigor**, **Credibilidade** e **Relevância**.

Cada critério foi operacionalizado por questões objetivas, totalizando nove itens, avaliados em uma escala ordinal de três níveis: **1** (atendido), **0,5** (parcialmente atendido) e **0** (não atendido). A pontuação final de cada estudo correspondeu à soma das pontuações atribuídas às questões, resultando em valores no intervalo de 0 a 9.

Os escores de qualidade foram utilizados como apoio na fase de análise e síntese dos dados, permitindo considerar a confiabilidade metodológica e a relevância dos estudos em relação aos objetivos da RSL.

[1] https://parsif.al/

4 Resultados

A estratégia de busca retornou inicialmente 623 artigos. Após a triagem por título e resumo, 512 estudos foram excluídos por não atenderem aos critérios definidos no protocolo da revisão, resultando em 111 artigos selecionados para leitura completa. Ao final da aplicação dos critérios de inclusão, exclusão e da análise detalhada do conteúdo, 60 artigos foram considerados elegíveis para extração de dados, compondo o corpus final desta Revisão Sistemática da Literatura.

4.1 Visão Geral dos Estudos Selecionados

A distribuição temporal dos estudos, indica uma concentração expressiva de publicações nos anos mais recentes. O maior volume ocorre em 2024, com 27 estudos, seguido por 2023 (13 estudos) e 2025 (11 estudos). Os anos iniciais do recorte temporal apresentam menor representatividade, com 6 estudos em 2022 e 3 em 2021, evidenciando que a intensificação da pesquisa acompanha a disseminação e o uso ampliado de LLM em contextos socialmente sensíveis.

O conjunto completo dos 60 artigos incluídos, identificados por um ID único (A01–A60) utilizado ao longo deste texto para referência cruzada, bem como seus respectivos Índices de Qualidade (IQ), encontra-se disponibilizado como material externo. De forma geral, os estudos apresentaram alta qualidade metodológica, com valores de IQ variando entre 8,0 e 9,0. [2]

A distribuição dos estudos por macro-categorias de modelos avaliados indica uma predominância de trabalhos que analisam modelos do tipo *encoder-based* e modelos de código aberto, ambos contemplados em 32 estudos. Em seguida, modelos proprietários aparecem em 18 estudos, refletindo o interesse da literatura em avaliar sistemas amplamente utilizados, apesar de suas restrições de acesso. Um número menor de trabalhos investiga modelos explicitamente orientados à equidade (*fairness-aware models*), totalizando 7 estudos, enquanto apenas 4 estudos analisam pipelines híbridos que combinam diferentes arquiteturas ou estratégias de processamento[3].

[2] As referências estão disponíveis no material externo.

[3] Um mesmo estudo pode abranger múltiplas categorias, uma vez que diversos trabalhos comparam ou avaliam mais de um tipo de modelo.

4.2 Estratégias de Mitigação de Viés em Modelos de Linguagem (RQ1)

As estratégias de mitigação identificadas nos estudos selecionados foram organizadas em cinco macro-categorias, conforme o estágio de aplicação da intervenção: (A) pré-processamento dos dados, (B) intervenções durante o treinamento, (C) estratégias em tempo de inferência, (D) pós-processamento das saídas e (E) ausência de mitigação. Essa organização, inspirada na taxonomia de Gallegos et al. (2024), adota um nível mais alto de abstração, consolidando subcategorias técnicas para viabilizar uma análise quantitativa e comparativa das abordagens reportadas na literatura. Adotamos o termo *inference-time* para as intervenções classificadas como *intra-processing* na taxonomia original e incluímos explicitamente uma categoria para estudos focados apenas em diagnóstico, sem aplicação de técnicas de mitigação. A distribuição das estratégias por macro-categoria é apresentada na Tabela 1.

A categoria **Pré-processamento** reúne intervenções nos dados antes do treinamento, como *data augmentation*, balanceamento e curadoria de amostras, incluindo abordagens baseadas em dados contrafactuais, substituição ou neutralização de termos sensíveis e geração sintética de dados mais equitativos. As estratégias de **In-training** atuam durante o treinamento ou *fine-tuning*, modificando parâmetros, funções de perda ou representações internas, com destaque para métodos adversariais, regularizações sensíveis à justiça, aprendizado de representações e adaptações eficientes de parâmetros.

As estratégias de **Inference-time** não alteram permanentemente os pesos do modelo e concentram-se na manipulação do prompt ou do contexto durante a inferência, incluindo *prompting* estruturado, *self-debiasing*, recuperação de contexto (RAG) e arquiteturas multiagente. Já as abordagens de **Pós-processamento** aplicam ajustes *post-hoc* sobre as saídas do modelo, como reranking, calibração de probabilidades ou poda de componentes responsáveis por comportamentos enviesados.

Por fim, a categoria **Não se aplica** é a mais frequente, e reúne estudos dedicados exclusivamente à detecção e mensuração de viés, sem avaliação de estratégias de mitigação, reiterando a necessidade exposta por (Brown et al., 2020).

4.3 Métricas e Métodos de Avaliação de Viés (RQ2)

A avaliação de viés em modelos de linguagem envolve múltiplas dimensões técnicas, normativas e metodológicas. Seguindo levantamentos recentes em justiça algorítmica e PLN (Mehrabi et al., 2021; Gallegos et al., 2024), as métricas e métodos identificados nesta revisão foram organizados segundo três dimensões analíticas: (i) o nível de acesso técnico ao modelo, indicando o que é efetivamente medido; (ii) a noção teórica de justiça operacionalizada; e (iii) a estrutura metodológica do procedimento de avaliação. Essa organização permite analisar não apenas quais métricas são utilizadas, mas também o tipo de viés que capturam e como são empiricamente aplicadas.

4.3.1 Nível de Acesso Técnico

Nesta dimensão, as métricas foram classificadas de acordo com o tipo de informação utilizada para quantificar o viés: (a) representações internas (*embeddings*), (b) distribuições de probabilidade e (c) texto final gerado. Essa distinção reflete diferentes graus de observabilidade do comportamento do modelo e diferencia avaliações de caráter mais diagnóstico daquelas orientadas ao impacto final da geração.

As métricas baseadas em *embeddings* avaliam o viés diretamente no espaço latente do modelo, partindo do pressuposto de que estereótipos sociais se manifestam como associações geométricas indevidas. Exemplos clássicos incluem o *Word Embedding Association Test* (WEAT) (Caliskan et al., 2017) e extensões para representações contextuais, como o SEAT, amplamente utilizadas para identificar viés representacional antes da aplicação em tarefas finais.

As métricas baseadas em *probabilidade* exploram preferências sistemáticas do modelo por sentenças estereotipadas em relação a alternativas neutras ou antiestereotipadas, por meio de tarefas de preenchimento de lacunas ou estimativas de *Pseudo-Log-Likelihood*. Benchmarks como *CrowS-Pairs* (Nangia et al., 2020) e *StereoSet* (Nadeem et al., 2021) operam nesse nível, frequentemente combinando medidas de viés e qualidade linguística.

Por fim, métricas baseadas em *texto gerado* analisam exclusivamente as saídas produzidas pelo modelo, sendo especialmente adequadas para cenários de acesso restrito ou de *caixa preta*. Essas abordagens incluem análises lexicais, o uso de clas-

Tabela 1: Distribuição das estratégias de mitigação de viés por macro-categoria de intervenção

Macro-Categoria	IDs dos Artigos	Contagem
A) Pré-processamento	A09, A16, A17, A23, A28, A32, A35, A39, A45, A49, A50, A51, A59	13
B) In-training	A02, A03, A07, A16, A18, A22, A23, A28, A35, A39, A51, A53, A54, A57, A59	15
C) Inference-time	A11, A12, A20, A26, A29, A30, A33, A38, A41, A48, A49, A56	12
D) Pós-processamento	A14, A31, A42	3
E) Não se aplica	A01, A04, A05, A06, A08, A10, A13, A15, A19, A21, A24, A25, A27, A34, A36, A37, A40, A43, A44, A46, A47, A52, A55, A58, A60	25

sificadores auxiliares (por exemplo, toxicidade ou sentimento) e métricas baseadas em léxicos normativos, como *HONEST* (Nozza et al., 2021). Embora menos informativas sobre a origem interna do viés, essas métricas permitem avaliar diretamente o impacto social das respostas geradas.

A Tabela 2 evidencia a predominância de métricas baseadas em *Texto Gerado*, refletindo a tendência recente de avaliar LLM como sistemas de caixa-preta e de priorizar o impacto observável para o usuário final. Métricas baseadas em *Embeddings* e *Probabilidades* aparecem com maior frequência em estudos de caráter diagnóstico, sendo comum, em trabalhos metodologicamente mais robustos, a combinação de múltiplos níveis técnicos para investigar a propagação do viés das representações internas até a saída final do modelo.

4.3.2 Definição Teórica de Justiça

A segunda dimensão organiza as métricas segundo a noção normativa de justiça que orienta a avaliação, conectando procedimentos empíricos a fundamentos teóricos da justiça algorítmica (Hardt et al., 2016; Mehrabi et al., 2021). Nessa dimensão, os estudos foram classificados em duas categorias principais: (a) justiça de grupo e (b) justiça individual.

A **Justiça de Grupo** busca garantir tratamento equitativo, em nível agregado, entre grupos definidos por atributos protegidos. Métricas clássicas incluem Paridade Demográfica, *Equalized Odds* e *Equal Opportunity*, amplamente empregadas em tarefas de classificação, decisão automatizada e moderação de conteúdo. Já a **Justiça Individual** enfatiza que indivíduos semelhantes devem receber decisões semelhantes, sendo operacionalizada predominantemente por testes contrafactuais, nos quais apenas o atributo sensível é alterado, mantendo-se o restante da entrada constante.

A Tabela 3 evidencia que a vasta maioria dos estudos adota critérios de Justiça de Grupo, enquanto abordagens baseadas em Justiça Individual aparecem de forma substancialmente menos frequente e, em geral, combinadas a métricas de grupo. Esse resultado sugere que a literatura prioriza avaliações agregadas de equidade, mesmo reconhecendo que modelos podem satisfazer critérios de grupo e, ainda assim, produzir discriminações em casos individuais específicos.

4.3.3 Estrutura do Método de Avaliação

A terceira dimensão descreve a estrutura experimental utilizada para avaliar o viés, considerando a organização dos dados de teste e o protocolo de interação com o modelo, conforme sistematizado por Gallegos et al. (2024). Os métodos identificados foram organizados em três categorias: (a) entradas contrafactuais, (b) testes de associação e (c) prompts e geração aberta.

As abordagens **Contrafactuais** utilizam pares de entradas que diferem apenas no atributo sensível, permitindo isolar o efeito causal da identidade protegida sobre a saída do modelo. Esses métodos são amplamente empregados em avaliações baseadas em probabilidades ou preferências entre versões estereotipadas e antiestereotipadas. Os **Testes de Associação** quantificam ligações semânticas entre conceitos e atributos sociais, operando em níveis sentenciais ou discursivos, e são especialmente adequados para diagnosticar vieses representacionais latentes. Já as abordagens baseadas em **Prompts e Geração Aberta** avaliam diretamente o texto produzido pelo modelo a partir de prompts livres, sendo comuns em cenários de caixa-preta e mais próximas de contextos reais de uso.

A Tabela 4 mostra que métodos contrafactuais e

Tabela 2: Classificação das métricas segundo o nível técnico de acesso ao modelo

Categoria	IDs dos Artigos	Contagem
Embeddings	A02, A15, A17, A19, A20, A21, A24, A27, A28, A36, A45, A48, A49, A50, A51, A52	16
Probabilidades	A03, A06, A08, A13, A16, A17, A18, A19, A20, A30, A31, A33, A43, A44, A48, A49, A50, A51, A52, A53, A56, A57	22
Texto Gerado	A01, A04, A05, A07, A09, A10, A11, A12, A13, A14, A16, A17, A21, A22, A23, A25, A26, A27, A28, A29, A30, A31, A32, A34, A35, A37, A38, A39, A40, A41, A42, A43, A45, A46, A47, A49, A51, A54, A55, A57, A58, A59, A60	43

Tabela 3: Classificação dos estudos segundo a definição teórica de justiça adotada

Categoria	IDs dos Artigos	Contagem
Justiça de Grupo	A01, A02, A03, A04, A05, A06, A07, A08, A09, A10, A12, A13, A14, A15, A16, A17, A18, A19, A20, A21, A22, A23, A24, A25, A26, A27, A28, A29, A30, A31, A32, A33, A34, A35, A36, A37, A38, A40, A41, A42, A43, A44, A45, A46, A47, A48, A49, A50, A51, A52, A53, A54, A55, A56, A57, A58, A59, A60	58
Justiça Individual	A11, A17, A25, A39, A51, A58	6

de associação são amplamente utilizados em cenários controlados, enquanto avaliações baseadas em geração aberta predominam em estudos voltados ao impacto observável da linguagem gerada. Em conjunto, esses resultados indicam uma literatura que combina avaliações diagnósticas e análises orientadas ao uso real, refletindo diferentes compromissos entre controle experimental e validade ecológica.

4.4 Tipos de Discriminação Social (RQ3)

A análise dos 60 estudos selecionados revela uma concentração expressiva em um conjunto relativamente restrito de categorias de discriminação social. Conforme ilustrado na Figura 1, o viés de *gênero* é, de longe, o mais investigado, estando presente em 52 artigos. Esses trabalhos abordam diferentes manifestações de sexismo, incluindo estereótipos ocupacionais, associações de liderança e cuidado, desigualdade em autoria, além de vieses em tarefas de recomendação, classificação e geração de texto.

A segunda categoria mais recorrente é *raça e etnia*, identificada em 27 estudos. Esses trabalhos analisam desde racismo explícito até formas mais sutis de discriminação, como o uso de nomes próprios como proxies raciais, o silenciamento de vozes negras, estereótipos étnicos e vieses herdados de datasets históricos amplamente utilizados. Em seguida, a discriminação baseada em *religião* aparece em 21 artigos, com foco predominante em islamofobia, antissemitismo e vieses pró-cristãos em tarefas de associação semântica e geração de conteúdo.

Outras categorias relevantes incluem viés *ocupacional, de classe social e status socioeconômico* (21 estudos), *nacionalidade e região* (15), *idade* (13) e *orientação sexual* (9). Essas categorias costumam ser investigadas de forma combinada, evidenciando a natureza interseccional do viés algorítmico. Em contraste, formas de discriminação como *capacitismo*, *aparência física*, *viés linguístico/dialetal* e *viés político* aparecem de maneira substancialmente menos frequente, indicando lacunas importantes na literatura atual.

4.5 Idiomas Avaliados (RQ4)

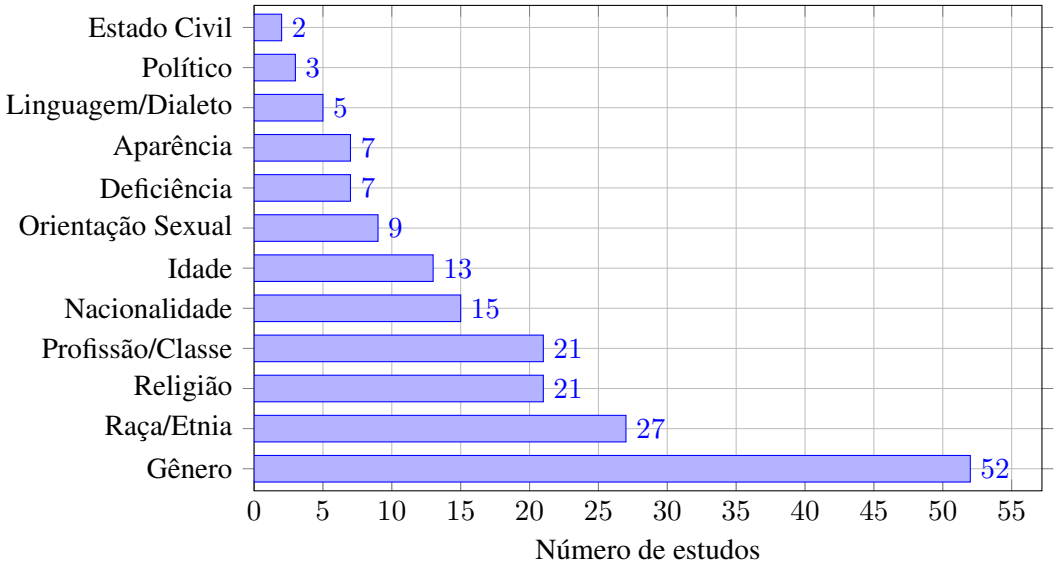
No que diz respeito aos idiomas em que as métricas de justiça e estratégias de mitigação foram avaliadas, observa-se uma dominância quase absoluta do inglês, como explicitado na Figura 2. Dos 60 estudos analisados, 57 conduzem suas avaliações exclusivamente em língua inglesa, refletindo tanto a disponibilidade de benchmarks padronizados quanto o foco histórico da área em modelos treinados majoritariamente nesse idioma.

A avaliação em outros idiomas é pontual e fragmentada. Apenas 9 estudos consideram os idiomas: bangla, norueguês, francês, mandarim, alemão, italiano e espanhol. Em geral, esses trabalhos surgem

Tabela 4: Método/Dataset de Avaliação dos Vieses

Categoria	IDs dos Artigos	Contagem
Contrafactual	A01, A03, A06, A07, A11, A13, A14, A17, A18, A19, A20, A25, A30, A31, A34, A35, A39, A41, A46, A48, A49, A50, A51, A52, A53, A56, A58	27
Associação	A02, A03, A04, A08, A15, A17, A18, A19, A20, A21, A24, A27, A28, A31, A33, A36, A45, A48, A49, A50, A52, A53, A55, A56, A57	25
Prompts / Geração	A04, A05, A09, A10, A12, A13, A14, A16, A21, A22, A23, A26, A28, A29, A30, A32, A37, A38, A40, A41, A42, A43, A44, A46, A47, A54, A55, A57, A58, A59, A60	31

Figura 1: Distribuição dos tipos de discriminação social abordados nos estudos analisados (RQ3)



em contextos específicos, como o uso de modelos regionais ou a adaptação de benchmarks para cenários linguísticos distintos.

Essa distribuição evidencia uma limitação estrutural da literatura: embora modelos de linguagem sejam cada vez mais utilizados em contextos multilíngues e globais, a maioria das métricas de viés, benchmarks e estratégias de mitigação permanece avaliada quase exclusivamente em inglês. Como consequência, há pouca evidência empírica sobre a eficácia dessas abordagens em idiomas com estruturas morfológicas, sintáticas e contextos socioculturais distintos, o que levanta questionamentos sobre a generalização dos resultados reportados.

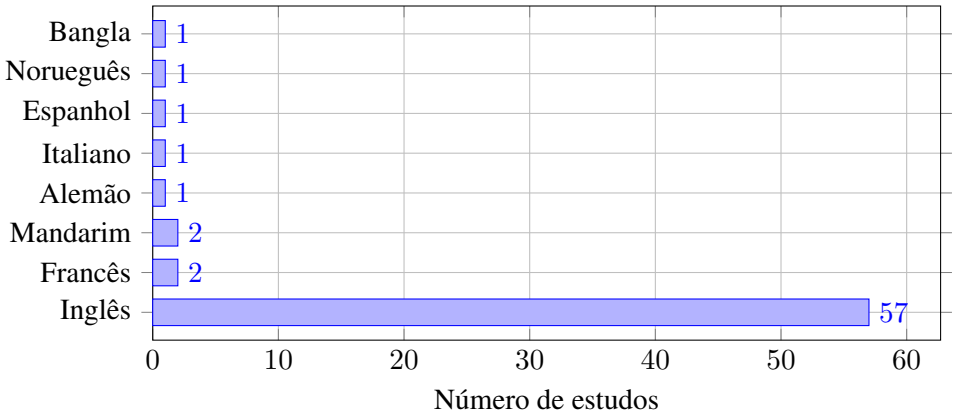
4.6 Principais Tendências, Limitações e Lacunas (RQ5)

A análise dos 60 estudos revela tendências consolidadas, bem como limitações metodológicas recorrentes e lacunas estruturais na literatura sobre viés em modelos de linguagem.

A principal limitação observada é a forte concentração das avaliações na língua inglesa, com escassa validação empírica em outros idiomas. Métricas, datasets e definições de estereótipos permanecem majoritariamente ancorados em contextos culturais anglófonos, o que levanta dúvidas quanto à generalização dos resultados para línguas de baixo recurso e contextos socioculturais distintos. Além disso, embora o viés de gênero seja amplamente investigado, ele é predominantemente tratado de forma binária, com baixa consideração de identidades não binárias e análises interseccionais envolvendo raça, idade, religião ou orientação sexual.

No plano metodológico, observa-se amplo uso de frases sintéticas, templates e testes de preenchimento de lacunas, que, apesar do controle experimental, apresentam baixa validade ecológica. O uso de léxicos estáticos também limita a captura de vieses implícitos e contextuais. Soma-se a isso um desequilíbrio entre estudos focados na detec-

Figura 2: Distribuição dos idiomas nos quais métricas de justiça e estratégias de mitigação foram avaliadas (RQ4)



ção de viés e aqueles que avaliam estratégias de mitigação, que, quando propostas, frequentemente sofrem com instabilidade, dependência de datasets específicos ou degradação de desempenho. Restrições de acesso a modelos proprietários e limitações computacionais também afetam a reprodutibilidade e a abrangência das avaliações.

Os estudos convergem na necessidade de ampliar avaliações para múltiplos idiomas, dialetos e contextos culturais, incluindo línguas de baixo recurso e cenários multilíngues. Outra direção recorrente é a incorporação explícita de interseccionalidade e identidades não binárias, superando análises baseadas em atributos isolados. No campo metodológico, destaca-se a demanda por avaliações com maior validade ecológica, utilizando dados do mundo real, feedback humano e participação de comunidades afetadas.

Há também consenso sobre a necessidade de amadurecer e padronizar métricas de justiça para LLM generativos, dado que muitas métricas clássicas não se adequam plenamente a modelos de grande escala. Por fim, os trabalhos apontam como agenda futura o desenvolvimento de estratégias de mitigação mais robustas e generalizáveis, bem como a expansão das análises para novas arquiteturas e aplicações de alto impacto social.

5 Discussão

Os resultados desta revisão evidenciam a predominância quase absoluta da língua inglesa nas avaliações de viés, o que reforça críticas de que métricas de justiça avaliadas exclusivamente nesse idioma apresentam limitações de generalização para outros contextos linguísticos e culturais (Blodgett et al., 2020; Gallegos et al., 2024). A escassez de estudos em línguas como português, espanhol e idiomas de baixo recurso sugere que vieses morfossintáticos e culturais específicos, como o gênero gramatical, permanecem subexplorados (Bender et al., 2021).

Observa-se também forte concentração no viés de gênero, geralmente tratado de forma binária, em detrimento de outras formas de discriminação e de análises interseccionais, o que limita a representatividade social das avaliações (Blodgett et al., 2020). Além disso, a literatura apresenta desequilíbrio entre diagnóstico e mitigação, com poucos estudos avaliando impactos práticos ou alocacionais das intervenções propostas.

Por fim, a predominância de métricas baseadas em texto gerado reflete a avaliação de LLM como sistemas de caixa-preta, especialmente em contextos de acesso restrito (Nadeem et al., 2021; Gallegos et al., 2024). Em conjunto, os achados indicam que, apesar dos avanços no diagnóstico de viés, a área ainda carece de avaliações multilíngues, interseccionais e orientadas ao uso real para o avanço da justiça algorítmica em modelos de linguagem.

6 Limitações do Trabalho

Apesar de seguir diretrizes consolidadas e protocolo explícito, esta RSL apresenta limitações. A busca restrita à base *Scopus* pode ter excluído estudos relevantes de outras fontes, e o recorte temporal entre 2020 e 2025, condicionado ao estado de indexação no momento da coleta, faz com que o corpus represente um retrato temporal do campo, possivelmente omitindo trabalhos recentes ainda não indexados e estudos anteriores com contribuições conceituais relevantes.

7 Conclusão

Este trabalho apresentou uma Revisão Sistemática da Literatura sobre viés, justiça algorítmica e es-

tratégias de mitigação em Grandes Modelos de Linguagem, analisando 60 estudos publicados entre 2020 e 2025. A revisão mapeou métricas de justiça, métodos de avaliação, estratégias de mitigação, idiomas analisados e tipos de discriminação social abordados pela literatura recente.

Os resultados evidenciam avanços na identificação de viés, sobretudo por meio de métricas baseadas em texto gerado, mas também revelam desequilíbrios persistentes, como a predominância de estudos em inglês, o foco no viés de gênero tratado de forma binária e a maior ênfase em diagnóstico do que em mitigação. Como contribuição, o trabalho destaca lacunas relacionadas ao multilinguismo, à interseccionalidade e à avaliação de impactos alocacionais, reforçando a necessidade de abordagens mais abrangentes e alinhadas aos contextos reais de uso de modelos de linguagem.

Uso de IA generativa

Ferramentas de IA generativa foram utilizadas exclusivamente para aprimoramento linguístico do texto (reescrita, parafraseamento e revisão), sem geração de novo conteúdo intelectual, em funções análogas a corretores gramaticais ou dicionários.

Referências

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, e Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? Em *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, página 610–623, New York, NY, USA. Association for Computing Machinery.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, e Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. Em *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, páginas 5454–5476, Online. Association for Computational Linguistics.

Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, e Adam T. Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. Em *Advances in Neural Information Processing Systems (NeurIPS)*, volume 29, páginas 4349–4357.

Pierre Bourdieu. 1991. *Language and Symbolic Power*. Harvard University Press, Cambridge, MA.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. Em *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Aylin Caliskan, Joanna J. Bryson, e Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

D. Carrera-Rivera, E. Y. Nakagawa, e S. de Faria Junior. 2022. Systematic literature review guidelines: Evolution and recent advances. *Journal of Systems and Software*, 192:111361.

Alessandro Fabris, Nina Baranowska, Matthew J. Dennis, David Graus, Philipp Hacker, Jorge Saldivar, Frederik Zuiderveen Borgesius, e Asia J. Biega. 2025. Fairness and bias in algorithmic hiring: A multidisciplinary survey. *ACM Trans. Intell. Syst. Technol.*, 16(1).

Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, e Nesreen K. Ahmed. 2024. Bias and fairness in large language models: A survey. *Computational Linguistics*, 50(3):1097–1179.

Moritz Hardt, Eric Price, e Nathan Srebro. 2016. Equality of opportunity in supervised learning. Em *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, página 3323–3331, Red Hook, NY, USA. Curran Associates Inc.

Barbara Kitchenham. 2007. Guidelines for performing systematic literature reviews in software engineering. *EBSE Technical Report*, EBSE-2007-01.

Michele Lamont, Graziella Moraes Silva, Jessica Welburn, Joshua Guetzkow, Nissim Mizrachi, Hanna Herzog, e Elisa Reis. 2016. Getting respect: Responding to stigma and discrimination in the united states, brazil, and israel. *Princeton University Press*.

Sharon Levy, Neha John, Ling Liu, Yogarshi Vyas, Jie Ma, Yoshinari Fujinuma, Miguel Ballesteros, Vittorio Castelli, e Dan Roth. 2023. Comparing biases and the impact of multilingual training across multiple languages. Em *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, páginas 10260–10280, Singapore. Association for Computational Linguistics.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, e Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM Comput. Surv.*, 54(6).

Moin Nadeem, Anna Bethke, e Siva Reddy. 2021. StereoSet: Measuring stereotypical bias in pretrained language models. Em *Proceedings of the 59th Annual Meeting of the Association for Computational*

Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), páginas 5356–5371, Online. Association for Computational Linguistics.

Nikita Nangia, Clara Vania, Rasika Bhalerao, e Samuel R. Bowman. 2020. CrowS-pairs: A challenge dataset for measuring social biases in masked language models. Em *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, páginas 1953–1967, Online. Association for Computational Linguistics.

Debora Nozza, Federico Bianchi, e Dirk Hovy. 2021. HONEST: Measuring hurtful sentence completion in language models. Em *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, páginas 2398–2406, Online. Association for Computational Linguistics.

Nripsuta Saxena, Karen Huang, Evan DeFilippis, Goran Radanovic, David C. Parkes, e Yang Liu. 2018. How do fairness definitions fare? examining public attitudes towards algorithmic definitions of fairness. *CoRR*, abs/1811.03654.

Emily Sheng, Kai-Wei Chang, Prem Natarajan, e Nanyun Peng. 2021. Societal biases in language generation: Progress and challenges. Em *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, páginas 4275–4293, Online. Association for Computational Linguistics.

Zeyu Tang, Jiji Zhang, e Kun Zhang. 2023. What-is and how-to for fairness in machine learning: A survey, reflection, and perspective. *ACM Comput. Surv.*, 55(13s).

Charles Tilly. 1998. *Durable Inequality*. University of California Press, Berkeley, CA.

Textual Inference in Portuguese: Comparing Language Models

Fabiana Avais[1,2], **Valeria de Paiva**[3], **and Livy Real**[4,5]
[1] Universidade Federal do Paraná
[2] Universidade Estadual de Ponta Grossa
[3] Topos Institute, Berkeley, USA
[4] Universidade Federal do Amazonas
[5] Instituto Kunumi
avaisfabiana@gmail.com valeria@topos.institute livy@kunumi.com

Abstract

Large language models (LLMs) are increasingly used for Natural Language Inference (NLI), yet their ability to perform logic-sensitive semantic reasoning, especially outside English, remains underexplored. This paper presents a preliminary investigation into the feasibility and usefulness of developing FraCaS-BR, a Portuguese adaptation of the FraCaS benchmark for semantic inference. Using a small diagnostic subset of seven FraCaS problems focusing on generalized quantifiers, plurals, and nominal anaphora, we evaluate the behavior of three LLMs (ChatGPT, Maritalk, and Evaristo) on Brazilian Portuguese translations. Each problem is submitted multiple times to assess correctness, variance, and consistency relative to the original FraCaS gold labels. The results reveal systematic differences across models. While ChatGPT shows higher overall correctness and stability, all models exhibit limitations that undermine their reliability on logic-controlled inference tasks. The extent of manual correction required during translation further underscores the necessity of human-in-the-loop evaluation. Taken together, these findings support and motivate the development of FraCaS-BR as a controlled evaluation resource for assessing semantic reasoning in Portuguese.

1 Introduction

The European Commission project *Framework for Computational Semantics*, abbreviated as FraCaS (Cooper et al., 1996)[1], ran from 1993 to 1996. Its goal was to develop an informal framework for comparing semantic approaches to language, both in terms of their theoretical claims and their suitability for implementation. The project was highly successful: the collection of examples devised to compare different formalisms is still in use today, more than thirty years later, as a benchmark for a wide range of reasoning and semantic tasks.

One of the main semantic tasks is still Natural Language Inference (NLI). There are plenty of resources, corpora, scripts, and competitions to evaluate NLI in English. However, there are fewer resources for NLI in Portuguese. The corpus ASSIN (for *Avaliação de Similaridade Semântica e Inferência Textual*) (Fonseca et al., 2016) is "a corpus annotated with pairs of sentences written in Portuguese that is suitable for the exploration of textual entailment and paraphrasing classifiers"[2]. The corpus does not consider contradictions, a serious issue if one thinks logically about inference.

A second corpus for NLI in Portuguese is SICK-BR [3], a careful translation of the SICK corpus of Marelli et al. (2014). Unlike ASSIN, the corpus SICK was meant to have simplified sentences, as far as linguistic phenomena are concerned. Thus, named entities, complicated time expressions, and world knowledge are kept at a minimum, as is the size of the vocabulary. The task, the corpus used for the task, and the evaluation results are described in Real et al. (2020).

InferBR (Bencke et al., 2024) is a third Portuguese NLI resource in which premises are semi-automatically generated, and hypotheses are subsequently generated automatically. Premises are constructed from two source datasets: PraCeGoVer (dos Santos et al., 2021) and SICK-BR (Real et al., 2018). These datasets are processed using GPT-4 and transformed into premise sentences. Hypotheses are then generated via few-shot prompt engineering, where each premise serves as input for producing three hypotheses corresponding to the labels entailment, contradiction, and neutral. The overall generation and evaluation process was reviewed by three human annotators.

[1] https://cordis.europa.eu/project/id/LRE62051

[2] https://huggingface.co/datasets/assin

[3] https://github.com/livyreal/SICK-BR

Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)
Vol. 2, pages 201–209
April 13-16, 2026

Our long-term goal is to produce a corpus called FraCas-BR (de Paiva and Real, 2024), a resource that goes back to the original goal of benchmarking a large selection of semantic phenomena. We hope to obtain manually checked translations of FraCaS sentences into Portuguese. We plan to verify, through careful annotation work, that the (mostly logical) phenomena described in English remain the focus of the new dataset and that the semantic phenomena 'behave' in Portuguese the same way they do in English. This is similar to the work in Amblard et al. (2020) for French and in the MultiFraCaS project[4] for Farsi, German, Greek, and Mandarin.

The original FraCas corpus consists of 346 'problems', each problem contains one or more premises and one question. There are a total of 536 premises, or an average of 1.55 premises per problem. This work presents a preliminary study on the feasibility and usefulness of developing FraCaS-BR. We conduct a small-scale experiment with seven examples, distributed across the three inference labels, entailment, contradiction and neutral. After translating these examples into Portuguese, we examine (i) whether the translations preserve the original logical relations and (ii) whether Portuguese-language LLMs can interpret them correctly. For this, we analyzed how three different LLMs (ChatGPT [5], Maritalk [6], and Evaristo [7]) deal with the FraCas problems in Portuguese. We conclude by summarizing our findings, which indicate a strong need for an evaluation resource of this kind for Portuguese.

Related Work

Haruta et al. (2020) proposed an end-to-end logic-based inference system for labeling both comparatives and generalized quantifiers, and evaluated it with FraCas. The system is successful in the task, and has five modules: implementation of a Combinatory Categorical Grammar parser, transformation to syntactic trees, semantic parsing, conversion to formal logics, and automatic inference.

Bernardy and Chatzikyriakidis (2021) created a similar NLI system, which transforms syntax trees into logical formulas (using the Rocq proof assistant[8]). Their goal was to handle temporal semantics in the whole FraCas dataset, and they obtained an overall accuracy of 81%, and 73% on temporal reference problems.

Amanaki et al. (2022) translated FraCas to Greek with human validation. They also added 428 new problems to the dataset, specifically to deal with Greek syntax.

Taken together, these studies underscore both the continued relevance of FraCaS as a benchmark for logical inference and the practicality of adapting it to new languages. However, little is known about how contemporary large language models perform on such tightly controlled, logic-oriented inference problems in Portuguese. In this preliminary study, we therefore examine seven FraCaS examples translated into Portuguese and analyze the responses produced by three large language models, offering a focused exploration of the challenges involved.

2 Evaluating Large Language Models

Evaluating the logical reasoning capabilities of LLMs is increasingly important. Although LLMs are now widely deployed and can generate fluent, persuasive text, evidence shows that they continue to struggle with basic logical reasoning tasks that are straightforward for humans (Suzgun et al., 2024). These models excel at predicting context and recognizing linguistic patterns, but this often comes at the expense of sound reasoning, with convincing chains of thought sometimes masking logical errors.

Recent work on legal applications (Trautmann et al., 2024) shows that combining LLMs with classical NLI frameworks yields strong performance in legal question answering, in part because NLI supports auditable claim verification. In this context, we argue that a general-purpose semantic resource for Portuguese would significantly strengthen the current evaluation landscape. Our preliminary study takes a first step in this direction by using a small, semantically and logically representative dataset to explore both the challenges of translating the FraCaS corpus into Portuguese and the logical reasoning behavior of LLMs in that language.

3 Our project

Given the current capabilities and widespread use of LLMs, a natural research question arises: is it still relevant to fully translate the FraCaS corpus into Portuguese, or can contemporary LLMs already handle its inference problems reliably? An-

[4] https://gu-clasp.github.io/multifracas/
[5] https://chat.openai.com
[6] https://www.maritaca.ai
[7] https://evaristo.ai
[8] https://rocq-prover.org/

swering this question requires evaluating not only whether LLMs assign the correct inference labels (correctness), but also how stable and confident their predictions are across runs (variance and consistency).

This leads to a secondary question: which LLM performs most effectively on semantic inference tasks in Brazilian Portuguese, and therefore produces the most reliable labels for FraCaS-style problems? By addressing these questions, we aim to determine whether a complete Portuguese translation of the FraCaS framework remains necessary, or whether its evaluative role is already subsumed by state-of-the-art LLMs.

FraCaS problems are organized by linguistic phenomenon, including generalized quantifiers, plurals, (nominal) anaphora, ellipsis, adjectives, comparatives, temporal reference, verbs, and attitudes. This experiment is intended both to probe LLM performance on logically controlled inference tasks in Portuguese and to inform best practices for translating the full FraCaS corpus based on observed model behavior.

3.1 Translating FraCas to Portuguese

The first task was to validate both the dataset's translation and its logical interpretation in Portuguese.

The FraCas sections selected for this work were generalized quantifiers, plurals, and nominal anaphora. More specifically, we chose seven problems listed in Table 1. A first experiment was conducted using the free versions of three large language models: ChatGPT, Maritalk, and Evaristo. These models were selected for complementary reasons. ChatGPT was included due to its widespread use and accessibility. Maritalk because it is a Brazilian LLM developed by Maritaca AI and trained specifically on Brazilian Portuguese data, with support for integration into platforms such as LangChain and Langflow. Evaristo, by contrast, is a recently released European Portuguese chatbot, designed around principles of open AI and user privacy, and built on open-source LLMs to promote transparency and community involvement. We use both variants of Portuguese (European and Brazilian) because criteria such as 'naturality' of the translation into Portuguese depends on the variant of Portuguese used by the human annotator.

The experimental runs took place in August 2025. Each of the seven problems was translated once by each model. Following a linguistic analysis of the outputs, we chose to manually correct the version produced by ChatGPT, as its initial translations more consistently preserved the intended semantic relations and exhibited a more natural syntactic structure, particularly in cases involving monotonicity. The final, curated translations are presented in Table 2.

3.2 Predicting NLI labels

After validating the translations and their logical adequacy, we turn to the second step of our investigation: evaluating how LLMs perform on the NLI task itself. Our experiment examines whether different models can correctly and consistently assign FraCaS-style inference labels to the translated problems, shifting the focus from translation quality to semantic reasoning behavior.

Here, we test the seven Portuguese problems across the three LLMs under investigation. Each problem instance was run ten times in each LLM, and the same prompt was used each time. The prompt was built with meta-prompting, which relies on writing a prompt and asking an LLM to improve it (Schulhoff et al., 2025). The prompt is in Portuguese below.

Analise se a HIPÓTESE decorre logicamente da PREMISSA. Responda apenas com uma das seguintes opções: YES, NO, ou UNKNOWN. Responda YES se a hipótese for uma consequência lógica da premissa (ou seja, sempre for verdadeira quando a premissa for verdadeira). Responda NO se a hipótese contradiz a premissa ou não pode ser verdadeira ao mesmo tempo que a premissa. Responda UNKNOWN se a premissa não fornece informação suficiente para determinar a veracidade da hipótese.[9]

This analysis focuses on a small subset of FraCaS problems involving generalized quantifiers (e.g. *todo, algum, poucos*). The subset is not intended to be statistically representative of the full FraCaS corpus; rather, it serves as an exploratory diagnostic sample. Because the FraCaS dataset spans a wide range of semantic phenomena, each section places different semantic demands on language models. Working with a limited subset therefore allows for an initial assessment of LLM be-

[9]Answer with only one of the following options: YES, NO, or UNKNOWN. Answer YES if the hypothesis is a logical consequence of the premise (that is, it is always true when the premise is true). Answer NO if the hypothesis contradicts the premise or cannot be true at the same time as the premise. Answer UNKNOWN if the premise does not provide enough information to determine the truth of the hypothesis.

ID	Premises	Hypothesis	Label
6	No really great tenors are modest.	Are there really great tenors who are modest?	NO
35	All Europeans can travel freely within Europe. Every European is a person. Every person who has the right to live in Europe can travel freely within Europe.	Do all Europeans have the right to live in Europe?	UNKNOWN
50	Every Canadian resident can travel freely within Europe. Every Canadian resident is a resident of the North American continent.	Can every resident of the North American continent travel freely within Europe?	UNKNOWN
96	The Ancient Greeks were all noted philosophers.	Was every Ancient Greek a noted philosopher?	YES
137	There are 100 companies. ICM is one of the companies and owns 150 computers. It does not have service contracts for any of its computers. Each of the other 99 companies owns one computer. They have service contracts for them.	Do most companies that own a computer have a service contract for it?	YES
211	All elephants are large animals. Dumbo is a small elephant.	Is Dumbo a small animal?	NO
223	The PC-6082 is faster than the ITEL-XZ. The PC-6082 is slow.	Is the ITEL-XZ fast?	NO

Table 1: Inference examples with premises, hypotheses, and labels

havior, making it possible to identify patterns in their predictions and gain insight into their general semantic competence.

Even with only seven problems, this diagnostic sample can reveal systematic behaviors, such as biases in label distribution or recurring inference errors. Moreover, failure to correctly handle these basic quantifier-related cases makes it unlikely that a model would perform well on the full FraCaS dataset. For this reason, a small but carefully chosen subset already provides meaningful evidence about model capabilities and limitations.

Using this subset also allows us to examine the stability of LLM responses and their ability to consistently capture semantic relations. Each problem was submitted ten times to each model, enabling analysis along three dimensions: correctness, variance, and consistency.

3.3 Comparing to Gold

In the first stage of the analysis, we evaluate the correctness of each model's predictions by aggregating results across the ten runs. The corresponding results are reported in Table 3.

For each problem, we determine the majority label across the ten runs and compare this aggregated prediction with the original FraCaS gold label. In this context, correctness is defined strictly as agreement with the FraCaS annotation. Table 3 summarizes these comparisons.

Overall, ChatGPT correctly labeled 6 out of the 7 problems, whereas Evaristo and Maritalk correctly labeled 4. When performance is broken down by label, ChatGPT correctly predicted 2 of the 2 'Yes' cases (IDs 96, 137), while Maritalk correctly predicted 1 and Evaristo predicted 1. For the 'No' label, ChatGPT and Evaristo achieved perfect performance (3/3), whereas Maritalk again labeled only 1 correctly. For the 'Unknown' label, Chat-

ID	Premises	Hypothesis	Label
6	Não há grandes tenores modestos.	Há grandes tenores modestos?	NO
35	Todos os europeus podem viajar livremente pela Europa. Todo europeu é uma pessoa. Toda pessoa que tem o direito de viver na Europa pode viajar livremente pela Europa.	Todos os europeus têm o direito de viver na Europa?	UNKNOWN
50	Todo residente canadense pode viajar livremente pela Europa. Todo residente canadense é um residente do continente norte-americano.	Todo residente do continente norte-americano pode viajar livremente pela Europa?	UNKNOWN
96	Os gregos antigos eram todos filósofos notáveis.	Todo grego antigo era filósofo notável?	YES
137	Existem 100 empresas. ICM é uma empresa e possui 150 computadores. Ela não tem contratos de serviço para nenhum dos seus computadores. Cada um das outras 99 empresas possui um computador. Elas têm contratos de serviço para eles.	A maioria das empresas que têm computador tem contratos de serviços?	YES
211	Todos os elefantes são animais grandes. Dumbo é um pequeno elefante.	Dumbo é um animal pequeno?	NO
223	O PC-6082 é mais rápido do que o ITEL-XZ. O PC-6082 é lento.	O ITEL-XZ é rápido?	NO

Table 2: Translated inference examples with premises, hypotheses, and labels

ID	Gold	ChatGPT	Maritalk	Evaristo
6	No	No	Unknown	No
96	Yes	Yes	Yes	No
35	Unknown	Unknown	Unknown	Yes
50	Unknown	No	Unknown	Yes
137	Yes	Yes	Unknown	Yes
211	No	No	No	No
223	No	No	Unknown	No

Table 3: Comparison between gold labels and model predictions.

GPT correctly predicted 1 case, Maritalk correctly predicted 2, and Evaristo did not predict this label correctly in any instance.

Taken together, the results in Table 3 indicate that ChatGPT exhibits higher overall correctness than both Evaristo and Maritalk, as well as a more balanced distribution of predicted labels. Maritalk shows a tendency to over-predict the Unknown label, while Evaristo fails to predict this label altogether.

3.4 Consistency of answers

The second stage of the analysis focuses on consistency across the ten runs for each model. Specifically, we measure how frequently a model assigns the same label to the same problem, which serves as an indicator of confidence in its predictions.

A first inspection of Table 4 shows that, for every problem, the models vary their predictions between at most two labels; no problem is assigned all three labels by any model. This suggests a baseline level of internal consistency in the models' responses.

A further pattern emerges from the distribution of label variation. Models rarely alternate between the logically contradictory labels Yes and No for the same problem. Instead, most variability involves pairs such as Yes/Unknown or No/Unknown. Instances of direct Yes/No alternation are rare: ChatGPT exhibits a single such case (ID 211), while Evaristo shows two (IDs 35 and 96). Moreover, the only case in which the aggregated major-

Table 4: ChatGPT, Maritalk and Evaristo predictions

ChatGPT											
ID	**Gold**	1	2	3	4	5	6	7	8	9	10
6	No	No	No	No	No	No	No	No	No	No	No
35	Unk	Unk	Yes	Unk	Unk	Yes	Unk	Yes	Unk	Unk	Yes
50	Unk	No	No	No	No	No	No	No	No	No	No
96	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
137	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
211	No	No	No	Yes	No	No	No	No	No	No	No
223	No	No	No	No	No	No	No	No	No	No	No
Maritalk											
ID	**Gold**	1	2	3	4	5	6	7	8	9	10
6	No	Unk	Unk	Unk	Unk	Unk	Unk	Unk	Unk	Unk	Unk
35	Unk	Unk	Unk	Unk	Unk	Unk	Unk	Unk	Unk	Unk	Unk
50	Unk	Unk	Unk	Unk	Unk	Unk	Unk	Unk	Unk	Unk	Unk
96	Yes	Yes	Yes	Unk	Unk	Yes	Yes	Unk	Yes	Unk	Unk
137	Yes	No	No	No	No	No	Unk	No	Unk	No	Unk
211	No	Unk	Unk	Unk	Unk	Unk	Unk	Unk	Unk	Unk	No
223	No	Unk	Unk	Unk	Unk	No	No	Unk	Unk	Unk	No
Evaristo											
ID	**Gold**	1	2	3	4	5	6	7	8	9	10
6	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
35	Unk	Yes	No	No	No	Yes	Yes	Yes	Yes	Yes	Yes
50	Unk	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
96	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	No	Yes
137	Yes	No	No	No	No	No	No	No	No	No	No
211	No	No	No	No	No	No	No	No	No	No	No
223	No	No	No	No	No	No	No	No	No	No	No

ity label conflicts as Yes/No arises with Evaristo (Table 1, ID 96). Overall, mislabeling involving Unknown is far more frequent than mislabeling between Yes and No, suggesting that the semantic interpretation of the Unknown label poses a particular challenge for LLMs. This is further reflected in systematic tendencies across models: Maritalk tends to over-predict Unknown, whereas Evaristo systematically under-predicts it.

Beyond label distribution, we analyze consistency across the ten runs for each model as an indicator of confidence in its predictions.

Consistency was evaluated at ID level, measuring the stability of the assigned labels across multiple model runs. For each label, only IDs that matched entirely the ten runs were considered. Under this metric, ChatGPT showed the highest consistency for the No label (0.75), moderate consistency for Yes (0.50), and no consistency for Unknown (0.00). Applying the same metric to Maritalk, a lack of consistency was observed for the Yes and No labels, since no ID that received these labels maintained the same label across the ten runs. In contrast, the Unknown label showed the moderate consistency of 0.43. Finally, Evaristo achieved moderate consistency with Yes (0.50) and No (0.60), and absence of the label Unknown. On this measure, ChatGPT emerges as the most consistent model.

Table 4 illustrates the variance in label distributions across models. ChatGPT shows the most balanced profile. Maritalk, by contrast, displays a strong bias toward 'Unknown', selecting this label too many times. This skewed distribution helps explain its lower overall correctness and highlights differing strategies adopted by LLMs when faced with semantic uncertainty.

Finally, we are at a loss to explain why Evaristo did not find a single 'Unknown' in the whole set. At the time of this research, Evaristo was in beta, so we expected this issue to evolve in subsequent versions. Also, Evaristo is designed to process Eu-

ropean Portuguese, but this focus alone is unlikely to explain the observed difficulty, since logical reasoning in both Portuguese variants is largely the same.

A natural follow-up question is whether LLMs appear more confident when their predicted label matches the gold standard. For ChatGPT, this does not seem to be the case: the model shows very little variation across runs, regardless of whether its aggregated prediction is correct or incorrect. More precisely, ChatGPT's predictions vary in only two (ID #35 and #211) of the seven problems, and it produces fewer mislabels overall than the other models.

For instance, in ID 35, where the correct label is Unknown, ChatGPT alternates between 'Unknown' (6) and 'Yes' (4) times. Despite this variance, the aggregated prediction is correct. A similar pattern appears in ID 211, where the correct label is 'No' and ChatGPT predicts 'No' in 9 out of 10 runs, with a single 'Yes'. In both cases, correctness is preserved at the aggregate level despite some instability. ChatGPT mislabels only one problem: ID 50 (predicted 'No' instead of Unknown) and it might be said that 'world knowledge' could have played a role.

Maritalk mislabels three problems (IDs 6, 137, and 223). In one of these cases (ID 6), there is no variation across runs: the model consistently assigns an incorrect label, yielding 100% incorrect predictions. In the remaining cases, variance is present. For ID 137, where the correct label is 'Yes', Maritalk predicts No (7 times) and Unknown (3 times). For ID 223, where the correct label is 'No', it predicts 'Unknown' in 7 runs and 'No' in 3. Even when Maritalk produces correct predictions, its answers often vary, but the more salient pattern is its systematic tendency to over-predict the 'Unknown' label, including in cases where 'Yes' or 'No' is warranted.

Evaristo mislabels the same number of problems as Maritalk (IDs 35, 50, and 96), but exhibits a different failure mode. The model never predicts the 'Unknown' label, and its outputs across the seven problems vary only between Yes and No. As a result, problems whose correct label is 'Unknown' are systematically misclassified, indicating a fundamental limitation for NLI tasks that rely on three-way inference distinctions.

In summary, ChatGPT outperforms the other models in both correctness and consistency, exhibiting the lowest variance and the most balanced label distribution. Maritalk's tendency to overuse 'Unknown' degrades its performance, while Evaristo's complete avoidance of this label makes it unsuitable for the NLI task considered here.

Taken together, these results highlight systematic and model-specific limitations in handling logic-sensitive semantic distinctions, particularly with respect to the Unknown label. The divergent behaviors observed suggest that current LLMs do not yet provide a reliable substitute for carefully curated inference Portuguese benchmarks. Rather than eliminating the need for resources such as FraCaS-BR, these findings reinforce their importance as controlled evaluation tools for diagnosing semantic competence in Portuguese. Moreover, the extent of manual correction required even in this small pilot study underscores the necessity of a human-in-the-loop approach to corpus construction and validation, especially for logic-based resources where subtle semantic errors can invalidate entire inference patterns. These considerations motivate the broader discussion of FraCaS-BR as both an evaluation benchmark and a methodological safeguard for assessing LLM reasoning in Portuguese.

4 Discussion

The results of this preliminary study point to clear and systematic limitations in current LLMs when confronted with logic-sensitive semantic inference tasks in Portuguese. Although all three models exhibit a baseline level of internal consistency, their behavior diverges sharply with respect to correctness, label balance, and the treatment of semantic underspecification. In particular, the 'Unknown' label emerges as a persistent source of difficulty, either being overused (Maritalk) or entirely avoided (Evaristo), with only ChatGPT showing a more balanced, though still imperfect, handling of three-way inference distinctions.

These findings suggest that LLMs do not uniformly internalize the semantic conditions underlying FraCaS-style inference. While models rarely oscillate between the logically contradictory labels 'Yes' and 'No', they frequently collapse uncertainty into 'Unknown' or eliminate it altogether. This pattern indicates that LLMs may rely on surface-level heuristics or distributional cues rather than robust semantic representations capable of sustaining underspecified or indeterminate interpretations.

Crucially, higher consistency does not always correlate with correctness. ChatGPT's performance

illustrates that stable predictions can still be wrong, while Maritalk's high consistency on incorrect labels reveals systematic semantic bias rather than random noise. These observations reinforce the need to evaluate LLMs along multiple dimensions (correctness, variance, and consistency) rather than accuracy alone. Logic-based benchmarks such as FraCaS are particularly well suited to exposing these distinctions, as small semantic errors can be clearly traced to specific linguistic phenomena.

Taken together, these results argue strongly for the relevance of FraCaS-BR as a dedicated evaluation resource for Portuguese. Rather than being obviated by state-of-the-art LLMs, the need for a carefully translated and validated FraCaS corpus becomes more pressing in light of their uneven performance. FraCaS-BR can serve both as a diagnostic benchmark for assessing LLM reasoning and as a methodological anchor for future work on semantic inference, hybrid symbolic–neural systems, and language-specific evaluation in Portuguese.

Finally, while this study is necessarily extremely limited in scale, its findings provide concrete guidance for future work. Expanding FraCaS-BR to cover additional semantic phenomena, increasing the number of annotated examples, and incorporating formal consistency metrics will allow for more robust comparisons across models and languages. More broadly, the results suggest that progress in LLM reasoning should be measured not only by fluency or task performance, but by the ability to respect the logical structure of meaning—a goal for which FraCaS remains a uniquely valuable benchmark.

5 Conclusions and Future Work

This work presented a preliminary evaluation of Large Language Models on a small subset of FraCaS problems translated into Brazilian Portuguese. Rather than aiming at broad generalization, the experiment was designed as a diagnostic analysis, intended to reveal whether systematic patterns of success and failure already emerge when LLMs are confronted with logically grounded inference problems in Portuguese.

The results indicate that, even in a limited setting, LLMs display difficulties with semantic inference as in the FraCaS framework. While all models performed relatively well on problems labeled 'No', performance varied considerably for 'Yes' and, most notably, for 'Unknown'. The latter emerged as the most challenging label for Maritalk and Evaristo models. This may suggest that some current LLMs do not reliably encode the distinction between the absence of information and logical contradiction, a distinction that is central to formal semantic reasoning.

Differences between models were not limited to overall correctness, but also involved systematic biases in label usage. ChatGPT achieved the highest correctness with relatively low variance across runs, indicating stable behavior across repeated queries. Maritalk showed a strong tendency to overlabel 'Unknown', which resulted in confident but often incorrect classifications for problems whose correct label were 'Yes' or 'No'. Evaristo[10], in contrast, seems to ignore the 'Unknown' label, reducing the task to a binary classification problem. These patterns suggest that model-specific strategies or training biases strongly influence how inference categories are interpreted.

From the perspective of resource development, these findings support the relevance of translating and adapting FraCaS to Portuguese. The observed errors cannot be attributed solely to translation issues or surface-level linguistic ambiguity, but instead reflect deeper challenges in modeling logical inference. As such, relying exclusively on state-of-the-art LLMs does not eliminate the need for carefully constructed semantic benchmarks. On the contrary, benchmarks like FraCaS-BR remain essential for diagnosing specific reasoning failures that are not easily captured by large-scale NLI datasets.

Finally, this study highlights the importance of evaluation frameworks grounded in semantics, especially in multilingual contexts. While LLMs have demonstrated impressive linguistic capabilities, their performance on logically controlled inference tasks remains uneven. Expanding FraCaS-BR and applying it to a broader set of models and semantic phenomena constitutes a natural next step toward a more principled evaluation of logical reasoning in Portuguese-language LLMs.

References

Eirini Amanaki, Jean-Philippe Bernardy, Stergios Chatzikyriakidis, Robin Cooper, Simon Dobnik,

[10]We ran the beta version of Evaristo, which may explain the limitations found in it. We reported the problematic cases to the developers, and that's why we want to test Evaristo in the future to see if it has improved, believing in the power of the community surrounding an open-source LLM.

Aram Karimi, Adam Ek, Eirini Chrysovalantou Giannikouri, Vasiliki Katsouli, Ilias Kolokousis, Eirini Chrysovalantou Mamatzaki, Dimitrios Papadakis, Olga Petrova, Erofili Psaltaki, Charikleia Soupiona, Effrosyni Skoulataki, and Christina Stefanidou. 2022. Fine-grained entailment: Resources for Greek NLI and precise entailment. In *Proceedings of the Workshop on Dataset Creation for Lower-Resourced Languages within the 13th Language Resources and Evaluation Conference*, pages 44–52, Marseille, France. European Language Resources Association.

Maxime Amblard, Clément Beysson, Philippe de Groote, Bruno Guillaume, and Sylvain Pogodalla. 2020. A french version of the fracas test suite. In *LREC 2020-Language Resources and Evaluation Conference*, page 9.

Luciana Bencke, Francielle Vasconcellos Pereira, Moniele Kunrath Santos, and Viviane Moreira. 2024. InferBR: A natural language inference dataset in Portuguese. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 9050–9060, Torino, Italia. ELRA and ICCL.

Jean-Philippe Bernardy and Stergios Chatzikyriakidis. 2021. Applied temporal analysis: A complete run of the FraCaS test suite. In *Proceedings of the 14th International Conference on Computational Semantics (IWCS)*, pages 11–20, Groningen, The Netherlands (online). Association for Computational Linguistics.

Robin Cooper, Dick Crouch, Jan van Eijck, Chris Fox, Josef van Genabith, Jan Jaspars, Hans Kamp, David Milward, Manfred Pinkal, Massimo Poesio, and Steve Pulman. 1996. Using the framework. the FraCaS Consortium.

Valeria de Paiva and Livy Real. 2024. Towards FraCas-BR. *OpenCor Workshop*.

Gabriel Oliveira dos Santos, Esther Luna Colombini, and Sandra Avila. 2021. #pracegover: A large dataset for image captioning in portuguese. *Preprint*, arXiv:2103.11474.

Erick Rocha Fonseca, Leandro Borges dos Santos, Marcelo Criscuolo, and Sandra Maria Aluísio. 2016. ASSIN: Avaliação de similaridade semântica e inferência textual. In *PROPOR*, pages 1–8.

Izumi Haruta, Koji Mineshima, and Daisuke Bekki. 2020. Logical inferences with comparatives and generalized quantifiers. *Preprint*, arXiv:2005.07954.

Marco Marelli, Luisa Bentivogli, Marco Baroni, Raffaella Bernardi, Stefano Menini, and Roberto Zamparelli. 2014. A SICK cure for the evaluation of compositional distributional semantic models. In *LREC 2014*.

Livy Real, Erick Fonseca, and Hugo Gonçalo Oliveira. 2020. The ASSIN 2 shared task: A quick overview. In *PROPOR 2020, Evora, Portugal*, volume 12037 of *LNCS*, pages 406–412. Springer.

Livy Real, Ana Rodrigues, Andressa Vieira e Silva, Beatriz Albiero, Bruno Guide, Bruna Thalenberg, Cindy Silva, Igor C. S. Câmara, Guilherme de Oliveira Lima, Rodrigo Souza, Milos Stanojevic, and Valeria de Paiva. 2018. SICK-BR: a portuguese corpus for inference. In *PROPOR 2018*.

Sander Schulhoff, Michael Ilie, Nishant Balepur, Konstantine Kahadze, Amanda Liu, Chenglei Si, Yinheng Li, Aayush Gupta, HyoJung Han, Sevien Schulhoff, Pranav Sandeep Dulepet, Saurav Vidyadhara, Dayeon Ki, Sweta Agrawal, Chau Pham, Gerson Kroiz, Feileen Li, Hudson Tao, Ashay Srivastava, and 12 others. 2025. The prompt report: A systematic survey of prompt engineering techniques. *Preprint*, arXiv:2406.06608.

Mirac Suzgun, Tayfun Gur, Federico Bianchi, Daniel E. Ho, Thomas Icard, Dan Jurafsky, and James Zou. 2024. Belief in the machine: Investigating epistemological blind spots of language models. *Preprint*, arXiv:2410.21195.

Dietrich Trautmann, Natalia Ostapuk, Quentin Grail, Adrian Pol, Guglielmo Bonifazi, Shang Gao, and Martin Gajek. 2024. Measuring the groundedness of legal question-answering systems. In *Proceedings of the Natural Legal Language Processing Workshop 2024*, pages 176–186, Miami, FL, USA. Association for Computational Linguistics.

Parsing Nheengatu: Performance Gains for a Brazilian Indigenous Universal Dependencies Treebank

Dominick Maia Alexandre and Leonel Figueiredo de Alencar
Universidade Federal do Ceará (UFC), Brazil
Av. da Universidade 2683 – 60.020-181 – Fortaleza – CE – Brazil
dominick@letras.ufc.br, leonel.de.alencar@ufc.br

Abstract

This paper evaluates the impact of expanding the UD_Nheengatu-CompLin treebank on parsing performance for Nheengatu, a Brazilian endangered Indigenous language. We hypothesized that the inclusion of annotated data would result in a 10% improvement in the Labeled Attachment Score (LAS). To test this hypothesis, we conducted a 10-fold cross-validation experiment using UDPipe 1.4.0 under two conditions: parsing with gold tokenization and gold tags, and automatic parsing from raw text. Statistical significance was determined using the Mann–Whitney U test. Although the expected gain was not achieved, the results show improvements in parsing accuracy and reduced variance across folds. The findings highlight the importance of corpus expansion and standardized annotation workflows for improving parsing performance in low-resource language scenarios and for supporting reproducible evaluation methods in the computational modeling of minority languages.

1 Introduction

Despite significant advances in Natural Language Processing (NLP) over the last decade, a substantial gap persists between high-resource and low-resource languages. This disparity is driven by the scarcity of annotated data for the latter and the concentration of technological development on the former (Joshi et al., 2020; Bird, 2020). In addition, many Indigenous languages present challenges for computational modeling, such as rich morphology, orthographic variation, and limited standardization, which hinder the direct application of existing NLP methods (Mager et al., 2018).

In this context, the Universal Dependencies (UD) project provides a shared cross-linguistic annotation framework that has been widely adopted for syntactic modeling across many well-resourced languages (Nivre et al., 2017; Church and Liberman, 2021), while also offering opportunities for extending NLP methods to low-resource ones (Thomas, 2019; Tyers and Henderson, 2021; Martín Rodríguez et al., 2022; da Silva and Pardo, 2024).

The Nheengatu treebank in the UD collection is a useful case study for examining the morphosyntactic annotation and parsing for low-resource languages. Once used as a *lingua franca* across the Amazon basin during the 18th and 19th centuries, Nheengatu is today an endangered language spoken primarily in Brazil. The UD_Nheengatu-CompLin treebank constitutes the first effort to provide a syntactically annotated corpus for the language following the UD framework (de Alencar, 2023; Alencar, 2024; de Alencar, 2024b,a, 2025).

Since its release in 2022, the treebank has been expanded and revised to improve coverage and consistency. In UD version 2.17, it received a 3.5-star rating, exceeding all other 24 Indigenous-language treebanks of the Americas ($\leq$ 2 stars) and approaching much larger treebanks of high-resource languages, such as UD_Portuguese-Porttinari (Duran et al., 2023) and UD_Portuguese-PetroGold (Souza et al., 2021) (4 stars).

Earlier parsing experiments on the UD_Nheengatu-CompLin were conducted by de Alencar (2024a), but they relied on a smaller and imbalanced dataset. Recent annotation efforts have incorporated historical data from Hartt's (1938), introducing a nineteenth-century Lower Amazon variety and enabling a reassessment of parsing performance under a more linguistically diverse dataset.

In this work, we evaluate parsing accuracy with and without the inclusion of Hartt's (1938), using an updated parsing pipeline and reproducible evaluation scripts. Although we hypothesized that this expansion would lead to an improvement of at least 10% in Labeled Attachment Score (LAS), the experiments revealed a more modest but still significant enhancement in parsing quality.

Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)
Vol. 2, pages 210–219

We conduct our experiments using UDPipe 1.4.0 (Straka et al., 2016), whose sensitivity to annotation quality makes it well suited for assessing the impact of treebank expansion in a low-resource setting.

The paper is organized as follows: Section 2 reviews related work on parsing for Brazilian and Amerindian languages within the UD framework; Section 3 presents Nheengatu and the linguistic phenomena recently included in the treebank; Section 4 describes the methodology; Section 5 reports the parsing results; Section 6 discusses the most frequent parser errors; and Section 7 concludes and outlines directions for future work.

2 Related work

Further improvements in the Labeled Attachment Score (LAS), the main metric for dependency parsing, are now limited less by model architecture than by the quality and size of the data (Lopes et al., 2024). This reliance on data contributes to disparities between high- and low-resource languages, with LAS values above 90% for languages such as English, Portuguese, and Russian, compared to below 40% for many minority languages.

Lopes and Pardo (2024) demonstrate that parsing accuracy degrades systematically as training data is reduced, with LAS dropping from 91.74% when trained on 5,893 sentences to 88.78% with only 1,473 sentences. These results show that even for well-resourced languages and gold-standard annotations, corpus size remains a decisive factor for parsing performance. Importantly, this effect becomes substantially more pronounced in low-resource settings, helping to explain the markedly lower LAS values reported by Vasquez et al. (2018) for Shipibo-Konibo and by Pugh et al. (2022) for Nahuatl. In such cases, limited training data reduces lexical and grammatical coverage, leading to severe constraints on parser generalization.

In the context of Indigenous languages and the Tupian language family, Blum's (2022) study shows that large multilingual models such as mBERT and RoBERTa perform poorly on Brazilian Indigenous languages. In zero-shot settings, accuracy rarely exceeds 40%. By contrast, models trained on closely related languages consistently achieve better results than those relying on typologically distant sources. These findings indicate that large multilingual models often fail to capture the grammatical structure of languages that are underrepresented in their training data (Blum, 2022).

Blum (2022) further shows that combining data from multiple related languages can outperform single-source models, even with small training sets. Their experiments indicate that as few as 50–60 annotated sentences can already yield measurable gains, particularly for PoS tagging. In contrast, dependency parsing remains more challenging, with lower transfer performance due to syntactic complexity (Blum, 2022).

3 The Nheengatu language

Once the most widely spoken language in the Brazilian Amazon, reaching parts of present-day Venezuela and Colombia, Nheengatu is now spoken by about 6,000 speakers in Brazil and faces challenges in intergenerational transmission (Navarro, 2012; Navarro et al., 2017; Eberhard et al., 2025).

Originating from Tupinambá and widely adopted by Jesuits, settlers, and different Indigenous groups, Nheengatu served social, political, and religious functions in Brazil during the seventeenth and eighteenth centuries. Despite a royal ban in the eighteenth century, it continued to be spoken into the early twentieth century (Borges, 1996; Rodrigues, 1986; Moore, 2014; Stradelli, 2014).

Today, Nheengatu is mainly spoken in the Upper Rio Negro region and remains an important marker of Amazonian identity. The language has been documented since the nineteenth century through grammars, religious texts, and literary works, which form a key part of the textual sources used in the UD_Nheengatu-CompLin treebank.

3.1 The UD_Nheengatu-CompLin treebank

Since its release in 2022, the UD_Nheengatu-CompLin treebank has been under continuous development. In the UD version 2.17 release (November 15, 2025), it underwent a substantial expansion in size and linguistic coverage, reaching 2,742 trees and 26,033 words. More than 600 sentences were added from nineteenth-century Nheengatu documented during Charles Frederick Hartt's Morgan Expedition (1870–1871) in the Lower Amazon region (Hartt, 1872, 1938). All new sentences were annotated according to the project guidelines and reviewed by a second annotator. This addition made Hartt's (1938) the second largest source in the treebank, after Avila's (2021), and the largest source of spoken-genre data. The updated distribution of sources is shown in Figure 1.

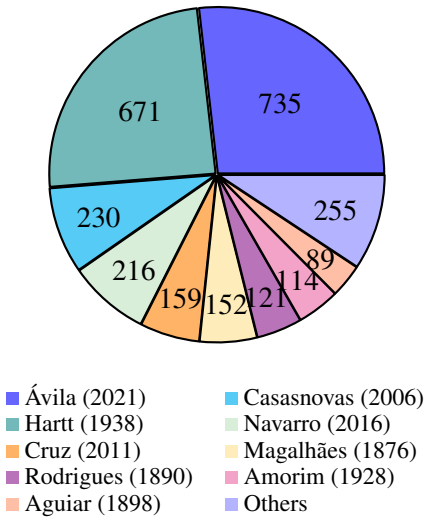


Figure 1: Sentence sources in UD_Nheengatu-CompLin ($n = 2742$). Sources with fewer than 89 sentences are grouped as *Others*.

(1) *E-kũi* *ne* *kiwira* *senúi*
2SG.IMP-go your brother [INF]call
u-ruri *arama* *ií* *ixéu.*
3SG.ACT-bring to water 1SG.DAT
'Go get your brother to bring me water.' (Hartt, 1938, p. 335)

As a result of this expansion, the treebank now includes a broader range of linguistic phenomena, such as archaisms, additional lexical forms, and morphosyntactic patterns not previously represented in the treebank (Alexandre and de Alencar, 2025). Example (1) illustrates morphosyntactic patterns no longer attested in contemporary Nheengatu. In present-day usage, imperative meanings are expressed using forms identical to the indicative. In contrast, the historical construction employs the auxiliary verb *ekũi* 'go' with archaic imperative morphology, including the second-person singular prefix *e-* and an irregular imperative form derived from the verb *sú* 'to go'.

The example also exhibits an SOV constituent order, with the full nominal object *ne kiwira* 'your brother' preceding the bare verb root. This indicates that historical Nheengatu allowed SOV order with full noun phrases, unlike contemporary Nheengatu, which displays a stable SVO pattern, likely influenced by long-term contact with Portuguese (da Cruz, 2011). From a typological perspective, however, the SOV order is more compatible with the language's postpositional system (Aikhenvald and Dixon, 2001).

Another archaic feature illustrated by the example is the absence of the double subject agreement that is typical of auxiliary constructions with verbs like *sú* 'to go' in contemporary Nheengatu. The sentence also contains the first-person dative pronoun *ixéu* 'to me', an inherited form from Old Tupi that had largely been replaced by the postposition *arama* by the time of Hartt's documentation (Avila, 2021), yet was still attested in nineteenth-century Lower Amazon Nheengatu.

Figure 2 displays the dependency graph for (1), illustrating how these historical morphosyntactic properties are encoded in UD_Nheengatu-CompLin. The annotation of Hartt's (1938) expanded the treebank with morphosyntactic patterns that are absent from contemporary data. These phenomena can now be analyzed using Yauti (de Alencar, 2023, 2025), an automatic annotation tool for Nheengatu, extending the treebank's linguistic coverage and supporting further computational analysis.

4 Methods

4.1 Parsing Experiments

Two parsing experiments were conducted using a 10-fold cross-validation procedure, following the methodology adopted in previous work on the UD_Nheengatu-CompLin treebank (de Alencar, 2024a). Parsing performance was evaluated using the development version of the treebank before and after the inclusion of historical material from Hartt's (1938).

The experiments were conducted using UDPipe 1.4.0, a trainable, language-agnostic pipeline for tokenization, morphosyntactic tagging, lemmatization, and dependency parsing of CoNLL-U data, which provides pre-trained models for Universal Dependencies treebanks and supports both gold-standard and fully automatic processing pipelines (Straka et al., 2016).

In **Experiment 1**, the parser was trained and evaluated on a version of the treebank excluding all 19th-century Lower Amazon data from Hartt (1938) (`nohartt.conllu`, 2,091 sentences), whereas **Experiment 2** used an expanded version including 743 sentences from Hartt (1938) (`all.conllu`, 2,834 sentences).[1]

[1]All data are available at: https://github.com/CompLin/nheengatu.

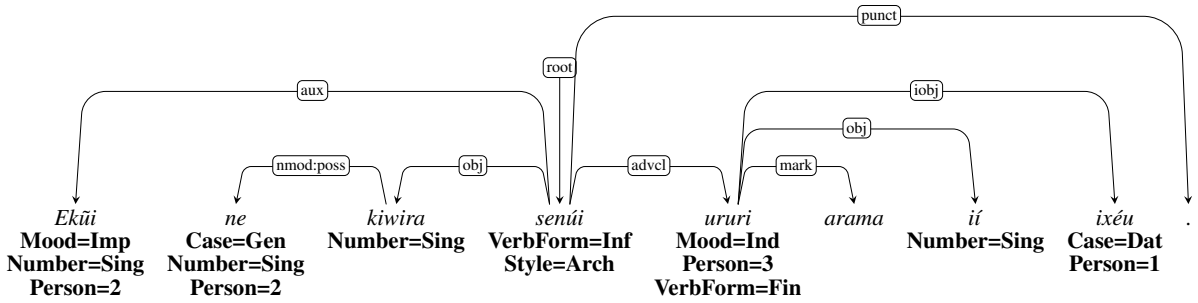


Figure 2: Dependency tree for (1).

For both experiments, the data were partitioned into 10 non-overlapping subsets of sentences of approximately equal size. In each fold, nine subsets were used for training and one for testing, such that every file and its sentences were used exactly once, as illustrated in Figure 3. This setup reduces dependence on a single train–test split and is well-suited for low-resource treebanks.

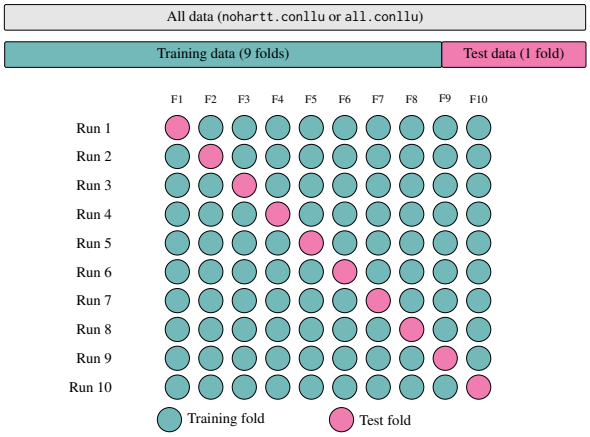


Figure 3: Illustration of 10-fold cross-validation. Each run uses one fold as test data (pink) and the remaining nine folds as training data (green).

Parsing performance was evaluated using the metrics **Labeled Attachment Score (LAS)** and **Unlabeled Attachment Score (UAS)**. Two evaluation conditions were considered: (i) parsing with gold tokenization and gold morphosyntactic annotations, and (ii) fully automatic parsing from raw text, in which UDPipe performed tokenization, tagging, and dependency parsing.

4.2 Parsing Performance Comparison

To assess the statistical significance of parsing performance differences between experimental settings, we used a Python script that reads per-fold LAS values directly from the parser output files and compares the distributions of LAS scores obtained under different experiments.

Statistical significance was evaluated using the **Mann–Whitney U test**, a nonparametric test for comparing two independent samples based on rank ordering (Mann and Whitney, 1947). For each comparison, a two-sided Mann–Whitney U test was performed on the two sets of LAS values.

The test was implemented using the `mannwhitneyu` function from the SciPy scientific computing library (Virtanen et al., 2020). In addition to reporting the U statistic and the corresponding two-tailed p-value, the script generates a fold-wise plot of LAS scores for visual inspection of performance differences across experiments.

The `CompareParsingResults.py` script was used to compare the ten fold-level LAS scores from Experiments 1 and 2, as well as LAS scores reported in prior work on the same treebank (de Alencar, 2024a). Because the earlier study used UDPipe 1.2, while our experiments used UDPipe 1.4.0, we performed two comparisons: one against the original UDPipe 1.2 results and one against a re-run of the experiment reported by (de Alencar, 2024a) using UDPipe 1.4.0.[2]

4.3 Error analysis

To identify the most frequent parsing errors, we analyze the outputs of 10 test splits under gold tokenization and tagging. For each split, gold and predicted CoNLL-U files were aligned and compared. We then compute a confusion matrix over UD dependency relations using a custom Python script (`deprel_confusion.py`), capturing label confusions independently of head assignment.

We further perform a diagnostic analysis that reports UAS and label accuracy and classifies errors into attachment-only, label-only, and combined

[2]Due to time constraints, we used the same models trained by de Alencar (2024a) for parsing with UDPipe 1.4.0.

	Exp. 1 (%)		Exp. 2 (%)	
Metric	**UAS**	**LAS**	**UAS**	**LAS**
Mean	87.06	82.70	88.12	84.29
SD	1.14	1.39	0.90	1.13

Table 1: Mean and standard deviation (SD) of parsing performance with **gold input** for Experiments 1 and 2.

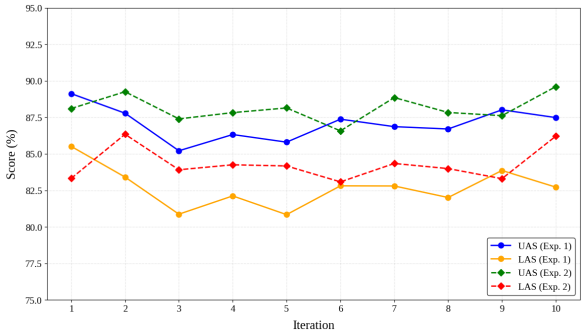


Figure 4: UAS and LAS with **gold input** across 10 folds for Experiments 1 and 2.

types. The most frequent errors are discussed in the following section.

5 Parsing Experiment Results

This section reports the results of the parsing experiments comparing two versions of the UD_Nheengatu-CompLin treebank. Although the initial hypothesis of a 10% improvement in parsing performance was not supported, the results indicate a positive effect of increased corpus size and coverage on parsing accuracy.

5.1 Parsing with Gold Input

Table 1 reports the mean and standard deviation of UAS and LAS for the two experiments with gold tokenization and gold tags. Across both metrics, Experiment 2 achieves higher mean scores than Experiment 1, with gains of 1.06 points in UAS and 1.59 points in LAS.

Experiment 2 also shows lower variance across folds for both metrics, indicating more stable performance under cross-validation (UAS SD = 0.90, LAS SD = 1.13) compared to Experiment 1 (UAS SD = 1.14, LAS SD = 1.39). Figure 4 shows the distribution of UAS and LAS scores across the ten iterations for both experiments.

5.2 Tokenization from Raw Text

We next evaluate tokenization performance under a fully automated parsing setting in which test sentences are provided as raw text, allowing the parser to perform tokenization. This setting allows us to assess the effects of corpus expansion when errors from earlier processing stages are not controlled.

	Exp. 1 (%)		Exp. 2 (%)	
Metric	**F1**	**SD**	**F1**	**SD**
Tokens	94.64	0.83	94.24	0.72
Multiword tokens	85.44	5.35	87.70	3.27
Words	94.42	0.80	94.09	0.69
Sentences	59.35	6.15	62.55	4.02

Table 2: Mean F1-score and standard deviation (SD) of tokenizer performance from **raw text** for Experiments 1 and 2.

	Exp. 1 (%)		Exp. 2 (%)	
Metric	**F1**	**SD**	**F1**	**SD**
UPOS tags	90.01	0.84	89.52	0.87
XPOS tags	89.18	0.87	88.73	0.91
Features	86.77	1.03	85.97	1.04
Lemmas	91.47	0.98	90.89	1.01

Table 3: Mean F1-score and standard deviation (SD) of tagging performance from **raw text** for Experiments 1 and 2.

Table 2 reports mean F1-scores and standard deviations for tokenization across Experiments 1 and 2, evaluated on tokens, multiword tokens, words, and sentences.

Token- and word-level segmentation remains high and stable across both experiments, with F1-scores above 94%. For multiword tokens, Experiment 2 achieves a higher mean F1-score than Experiment 1 (+2.26), along with a lower standard deviation across folds. Sentence segmentation remains the most challenging subtask in both experiments; however, Experiment 2 again shows higher mean performance (+3.20) and reduced variability.

5.3 Tagging from Raw Text

Table 3 reports tagging performance for UPOS, XPOS, morphological features, and lemmatization. Across all tagging components, Experiment 2 shows slightly lower mean scores than Experiment 1.

Within the UD framework, UPOS tagging relies on a smaller and more abstract label set than XPOS tagging and morphological feature prediction. In contrast, morphological features encode language-specific distinctions and combinations of features, which introduce additional complexity.

The larger differences observed for FEATS are

	Exp. 1 (%)		Exp. 2 (%)	
Metric	**UAS**	**LAS**	**UAS**	**LAS**
Mean	74.46	68.73	74.80	69.56
SD	1.96	2.17	0.95	0.91

Table 4: Mean and standard deviation (SD) of parsing performance from **raw text** for Experiments 1 and 2.

consistent with the increased linguistic and orthographic variability introduced in Experiment 2, which incorporates historical data with a wider range of morphosyntactic patterns, part-of-speech and feature combinations, and preserved orthographic variation from the source material.

5.4 Parsing from Raw Text

Table 4 reports UAS and LAS results for Experiments 1 and 2. As expected, performance is substantially lower than in the gold-input setting, reflecting the accumulation of errors from tokenization and morphosyntactic tagging. Experiment 2 shows small gains in mean UAS (+0.34) and LAS (+0.83), as well as reduced variance across folds, with standard deviations below 1% for both metrics compared to Experiment 1, which shows standard deviations close to 2%.

This pattern indicates a more stable parsing behavior under fully automatic conditions, even when errors from earlier processing stages may propagate to dependency parsing.

5.5 Statistical Significance

To assess whether the differences are statistically reliable, we apply a two-sided Mann–Whitney U test to the fold-level LAS scores. Under gold-input conditions, the test yields $U = 14.0$ and $p = 0.0073$, indicating a statistically significant difference between Experiments 1 and 2 (Figure 5).

We further compare Experiment 2 with the results reported by de Alencar (2024a), obtained using UDPipe version 1.2. As shown in the left panel of Figure 6, Experiment 2 consistently achieves higher LAS values across all folds when compared to the 2024 experiment. The observed differences are modest but systematic, generally ranging between approximately 1 and 3 percentage points, and are statistically significant ($p = 0.00032$).

Together with the fold-level patterns shown in Figure 5, this result suggests that the higher LAS scores observed for Experiment 2 are unlikely to be due to random variation across folds.

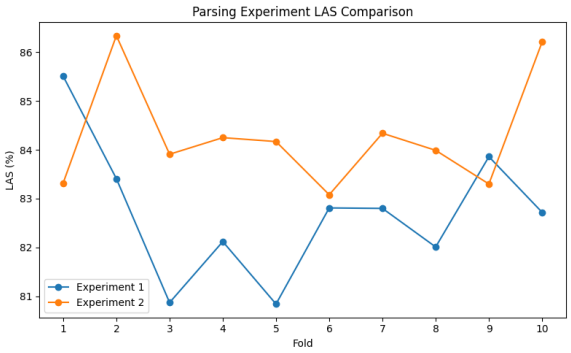


Figure 5: LAS comparison between Experiments 1 and 2.

However, this improvement primarily reflects the effects of corpus expansion, as the 2024 experiment was based on a smaller dataset (1,336 sentences). Contrary to initial expectations, differences in the parsing pipeline appear to play a minimal role, as re-running the earlier experiment with UDPipe 1.4.0 yields results that are essentially equivalent to those obtained with UDPipe 1.2.

As shown in the right panel of Figure 6, Experiment 2 consistently achieves higher LAS scores across all folds, and these differences remain stable when using UDPipe 1.4.0, remaining statistically significant (Mann–Whitney $U = 98.0$, $p = 0.00033$).

6 Frequent parsing errors

The most frequent confusion involves the core argument relations *nsubj* and *obj* (Figure 7). Across the test splits, 169 gold *nsubj* relations were predicted as *obj*, and 140 gold *obj* relations as *nsubj*. Figure 8 illustrates the former: the parser analyzes the subject *mbira-itá* ‘my children’ of *ukiri* ‘sleep’ as the object of the preceding subordinate verb *asika* ‘arrive’. Figure 9 shows the correct analysis.

The prediction in Figure 8 is not entirely implausible. In sentences with two verbal clauses, a noun phrase occurring between the verbs may, in principle, be interpreted either as the object of the first verb or as the subject of the second. In addition, Nheengatu is a pro-drop language, allowing the subject position to remain unexpressed, which makes an analysis in which *mbira-itá* does not attach to *ukiri* structurally possible. However, sentence (2) is not genuinely ambiguous, since the embedded verb is intransitive and does not license an object.

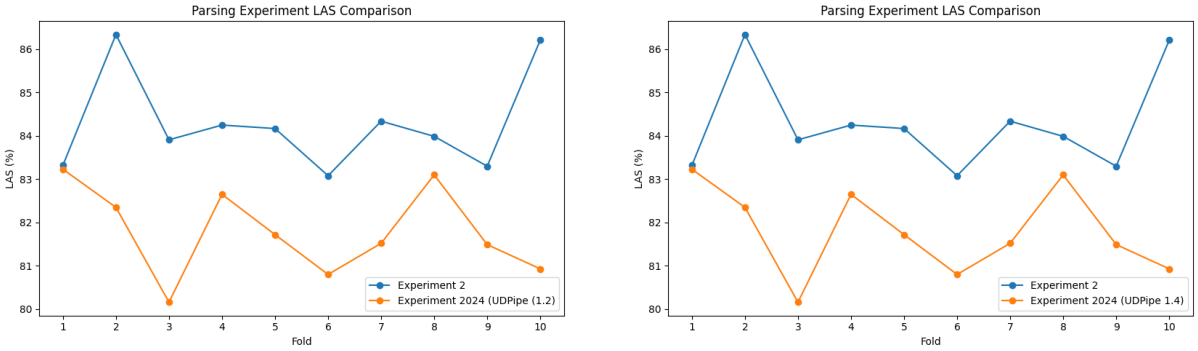


Figure 6: LAS comparison between Experiment 2 and the 2024 experiment using different UDPipe versions (v1.2 on the left, v1.4.0 on the right).

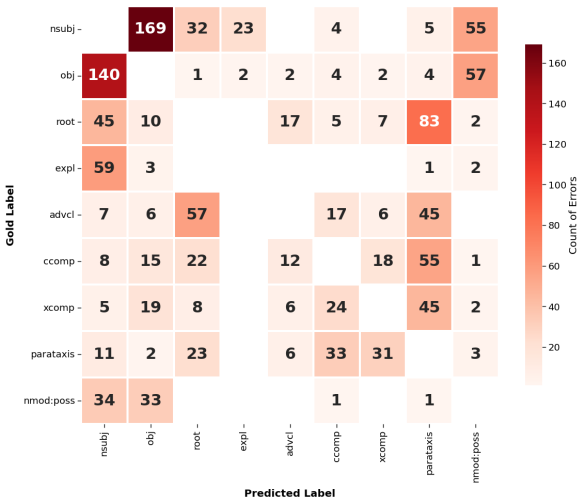


Figure 7: Heatmap of the most frequent UD Relations confusions (Raw Counts)

(2) *Asika* *ramé* *se* *mbira-itá*
1SG.arrive when 1SG.GEN child.PL
ukiri *ana* *uikú.*
3SG.sleep PFV 3SG.be

'When I arrive, my children are already sleeping.' (Moore et al., 1994, p. 110)

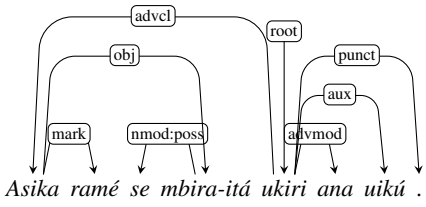


Figure 8: Incorrect dependency tree for (2).

While this error also occurs in shorter, typically monoclausal sentences, where the parser fails to recognize postverbal subjects (e.g., of unaccusative verbs), 78.2% of cases occur in sentences ranging from nine to 49 tokens. Longer sentences typically involve multiple coordinated, juxtaposed, or embedded predicates (e.g., in relative, adverbial, and complement clauses). Together, these patterns suggest that the parser struggles both to identify clause boundaries and to take verb valency into account.

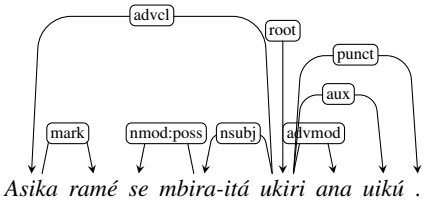


Figure 9: Gold dependency tree for (2).

The heatmap also reveals a frequent confusion between *root* and *parataxis*, reflecting the parser's difficulty in identifying the main predicate in complex sentences. Similarly, clausal relations such as *advcl*, *ccomp*, and *xcomp* are often predicted as *parataxis* or *root*. Table 5 summarizes the corresponding metrics and error distribution. Attachment-only errors are more frequent than label-only errors, suggesting that incorrect head assignment is a more common source of error than label misclassifications alone. In addition, a substantial number of cases (1,434) involve errors in both head and label, indicating that some errors arise in more complex sentences rather than from isolated misclassifications.

7 Summary of Results

Across evaluation settings, Experiment 2 yields higher parsing accuracy and reduced variability across folds. Expanding the UD_Nheengatu-CompLin treebank leads to consistent improve-

Metric	Value
Tokens compared	26,785
UAS	88.11%
Label accuracy	90.82%
LAS	84.28%
Attachment-only errors	1,750
Label-only errors	1,026
Head+label errors	1,434

Table 5: Global parsing performance and error distribution across all test splits.

ments in accuracy and stability. Error analysis indicates that errors concentrate in core arguments (e.g., *nsubj* vs. *obj*) and clause-level relations (e.g., *root*, *parataxis*, *advcl*).

Although the expected 10% gain was not observed, the results point to the importance of incremental treebank expansion combined with clear annotation guidelines, internal consistency, and review by a second annotator when developing resources for a low-resource language.

8 Final Remarks

This study investigated how the expansion of the UD_Nheengatu-CompLin treebank affects dependency parsing performance under controlled (gold tokenization and gold tags) and fully automatic conditions. By extending the treebank with historical nineteenth-century material, we assessed how increased data volume and linguistic diversity interact with different stages of a modern parsing pipeline for a low-resource language.

Across evaluation settings, the expanded treebank is associated with more stable parsing performance, reflected in lower variance across cross-validation folds. Under gold-input conditions, the differences between experiments are statistically significant, suggesting that increased annotated data supports more consistent syntactic predictions when tokenization and morphosyntactic annotation are fixed. In fully automatic parsing, average accuracy differences are small, but Experiment 2 shows lower variability across folds, indicating more stable model behavior.

The inclusion of historical data introduced greater linguistic heterogeneity, including orthographic variation and less frequent morphosyntactic patterns. While this diversity expanded syntactic coverage, it also increased the difficulty of morphosyntactic tagging. These findings underscore the importance of interpreting parsing results in relation to both corpus composition and processing conditions, particularly for low-resource Indigenous languages.

Some limitations of this study point to directions for future work. For instance, genre distribution in the treebank was not controlled or analyzed. A more detailed evaluation by genre would require prior classification of sentences, which is not currently available and is challenging due to the heterogeneous nature of the sources (e.g., narrative texts, grammatical examples, and mixed materials).

We plan to further expand the UD_Nheengatu-CompLin treebank, along with a more fine-grained analysis of dependency relation labels that remain challenging for the parsing pipeline, focusing on error patterns at the level of specific relations in order to identify systematic sources of ambiguity and inform refinements to annotation guidelines or modeling strategies. Another direction for future work is the evaluation of alternative parsing architectures.

Beyond Nheengatu, the methodology adopted in this study was designed to be reproducible, and the release of updated resources and evaluation scripts enables the same approach to be applied to other Indigenous and low-resource languages within the Universal Dependencies framework, supporting transparent resource development and comparable evaluation of parsing performance across minority languages.

Acknowledgments

This work was supported by the Brazilian CAPES Foundation and FAPESP (Grant No. 22/09158-5, DACILAT project at UNICAMP). We thank the anonymous reviewers for their helpful suggestions. We acknowledge the use of large language models (ChatGPT, Gemini, and Grammarly) for grammar and style revision, as well as for coding support.

References

Alexandra Y. Aikhenvald and R. M. W. Dixon. 2001. Introduction. In Alexandra Y. Aikhenvald and R. M. W. Dixon, editors, *Areal diffusion and genetic inheritance: Problems in comparative linguistics*, pages 1–26. Oxford University Press, Oxford.

Leonel Figueiredo de Alencar. 2024. UD_Nheengatu-CompLin: o corpus sintaticamente anotado do nheengatu da coleção Universal Dependencies. In *Anais*

Eletrônicos do XVI Encontro de Linguística de Corpus e da XII Escola Brasileira de Linguística Computacional, volume 1, pages 105–109, Brasília. Associação Brasileira de Linguística de Corpus.

Dominick Maia Alexandre and Leonel Figueiredo de Alencar. 2025. Universal Dependencies for 19th-Century Nheengatu from the Lower Amazon Region. In *Anais do XVI Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 588–598, Porto Alegre, RS, Brasil. SBC.

Marcel Twardowsky Avila. 2021. *Proposta de dicionário nheengatu-português*. Ph.D. thesis, Faculdade de Filosofia, Letras e Ciências Humanas da Universidade de São Paulo.

S. Bird. 2020. Decolonising speech and language technology. In *COLING 2020 - 28th International Conference on Computational Linguistics, Proceedings of the Conference*, COLING 2020 - 28th International Conference on Computational Linguistics, Proceedings of the Conference, pages 3504–3519. Association for Computational Linguistics (ACL).

Frederic Blum. 2022. Evaluating zero-shot transfers and multilingual models for dependency parsing and POS tagging within the low-resource language family tupían. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 1–9, Dublin, Ireland. Association for Computational Linguistics.

Luiz Carlos Borges. 1996. O nheengatú: uma língua amazônica. *Papia*, 4(2):44–55.

Kenneth Church and Mark Liberman. 2021. The future of computational linguistics: On beyond alchemy. *Frontiers in Artificial Intelligence*, 4:1–18.

Aline da Cruz. 2011. *Fonologia e gramática do nheengatú: A língua falada pelos povos Baré, Warekena e Baniwa*. LOT, Utrecht.

Diego Pedro Gonçalves da Silva and Thiago Alexandre Salgueiro Pardo. 2024. Grammar induction for Brazilian indigenous languages. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 2*, pages 64–72, Santiago de Compostela, Galicia/Spain. Association for Computational Lingustics.

Leonel Figueiredo de Alencar. 2023. Yauti: A tool for morphosyntactic analysis of Nheengatu within the Universal Dependencies framework. In *Anais do XIV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 135–145, Porto Alegre, RS, Brasil. SBC.

Leonel Figueiredo de Alencar. 2024a. A Universal Dependencies Treebank for Nheengatu. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese*, volume 2, pages 37–54, Santiago de Compostela, Galicia, Spain. Association for Computational Linguistics.

Leonel Figueiredo de Alencar. 2024b. Aspectos da construção de um corpus sintaticamente anotado do nheengatu no modelo Dependências Universais. *Texto Livre*, 17:e52653.

Leonel Figueiredo de Alencar. 2025. Enhancing a Nheengatu Morphosyntactic Analyzer for Word Formation and Non-standard Language. In *Anais do XVI Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 13–28, Porto Alegre, RS, Brasil. SBC.

Magali Duran, Lucelene Lopes, Maria das Graças Nunes, and Thiago Pardo. 2023. The Dawn of the Porttinari Multigenre Treebank: Introducing its Journalistic Portion. In *Anais do XIV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 115–124, Porto Alegre, RS, Brasil. SBC.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig, editors. 2025. *Ethnologue: Languages of the World*, 28 edition. SIL International, Dallas.

Charles Frederick Hartt. 1872. Notes on the Lingoa Geral or Modern Tupi of the Amazonas. *Transactions of the American Philological Association*, 3:58–76.

Charles Frederick Hartt. 1938. Notas sobre a língua geral, ou tupí moderno do Amazonas. *Anais da Biblioteca Nacional do Rio de Janeiro*, LI:305–390. [1929].

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The State and Fate of Linguistic Diversity and Inclusion in the NLP World. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Lucelene Lopes and Thiago Pardo. 2024. Towards Portparser - a highly accurate parsing system for Brazilian Portuguese following the Universal Dependencies framework. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 1*, pages 401–410, Santiago de Compostela, Galicia/Spain. Association for Computational Lingustics.

Lucelene Lopes, Thiago Pardo, and Magali Duran. 2024. Syntactic parsing: where are we going? In *Anais do XV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana (STIL 2024)*, pages 67–74, Porto Alegre, RS, Brasil. SBC.

Manuel Mager, Ximena Gutierrez-Vasques, Gerardo Sierra, and Ivan Meza-Ruiz. 2018. Challenges of language technologies for the indigenous languages of the Americas. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 55–69, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Henry B. Mann and Donald R. Whitney. 1947. On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *The Annals of Mathematical Statistics*, 18(1):50–60.

Lorena Martín Rodríguez and 1 others. 2022. Tupían language ressources: Data, tools, analyses. In *Proceedings of the 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages*, pages 48–58, Marseille, France. European Language Resources Association.

Denny Moore. 2014. Historical development of Nheengatu (Língua Geral Amazônica). In Salikoko S. Mufwene, editor, *Iberian Imperialism and Language Evolution in Latin America*, pages 108–142. University of Chicago Press, Chicago.

Denny Moore, Sidney Facundes, and Nádia Pires. 1994. Nheengatu (Língua Geral Amazônica), its history, and the effects of language contact. In *Proceedings of the Meeting of the Society for the Study of the Indigenous languages of the Americas, July 2-4, 1993 and the Hokan-Penutian Workshop, July 3, 1993*, pages 93–118, Berkeley, CA. [University of California]. Acesso em: 26 jul. 2024.

Eduardo de Almeida Navarro. 2012. O último refúgio da língua geral no Brasil. *Estudos Avançados*, 26(76):245–254.

Eduardo de Almeida Navarro, Marcel Twardowsky Ávila, and Rodrigo Godinho Trevisan. 2017. O Nheengatu, entre a vida e a morte: A tradução literária como possível instrumento de sua revitalização lexical. *Revista Letras Raras*, 6(2):9–29.

Joakim Nivre, Daniel Zeman, Filip Ginter, and Francis Tyers. 2017. Universal Dependencies. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Tutorial Abstracts*, Valencia, Spain. Association for Computational Linguistics.

Robert Pugh, Marivel Huerta Mendez, Mitsuya Sasaki, and Francis Tyers. 2022. Universal Dependencies for Western Sierra Puebla Nahuatl. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5011–5020, Marseille, France. European Language Resources Association.

Aryon Dall'Igna Rodrigues. 1986. *Línguas Brasileiras: Para o conhecimento das línguas indígenas*. Loyola, São Paulo. Vários quadros numerados e outros sem numeração.

Elvis Souza, Aline Silveira, Tatiana Cavalcanti, Maria Castro, and Claudia Freitas. 2021. PetroGold – Corpus padrão ouro para o dominio do petroleo. In *Proceedings of the 13th Brazilian Symposium in Information and Human Language Technology*, pages 29–38, Porto Alegre, Brazil. Association for Computational Linguistics.

Ermanno Stradelli. 2014. *Vocabulário português-nheengatu, nheengatu-português*. Ateliê Editorial, Cotia, SP. Original work published in 1929.

Milan Straka, Jan Hajič, and Jana Straková. 2016. UDPipe: Trainable pipeline for processing CoNLL-U files performing tokenization, morphological analysis, POS tagging and parsing. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4290–4297, Portorož, Slovenia. European Language Resources Association (ELRA).

Guillaume Thomas. 2019. Universal Dependencies for Mbyá Guaraní. In *Proceedings of the Third Workshop on Universal Dependencies (UDW, SyntaxFest 2019)*, pages 70–77, Paris, France. Association for Computational Linguistics.

Francis M. Tyers and Robert Henderson. 2021. A corpus of K'iche' annotated for morphosyntactic structure. In *Proceedings of the First Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP)*.

Alonso Vasquez and 1 others. 2018. Toward Universal Dependencies for Shipibo-konibo. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, pages 151–161, Brussels, Belgium. Association for Computational Linguistics.

Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, and 16 others. 2020. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272.

Bridging Cultural Gaps in Automated Translation of Brazilian Expressions: A Study on Cultural Adaptation

Maria Luiza Silva de Oliveira[1], Andressa Andrade Oliveira dos Santos[1]
Leandro Jose Silva Andrade[2]

[1]Federal University of Bahia (UFBA), Department of Executive Secretariat, Salvador, Brazil
[2]Federal University of Bahia (UFBA), Salvador, Brazil

mlso97@outlook.com, andressaandrade1810@gmail.com, leandrojsa@ufba.br

Abstract

Automated translation systems exhibit a tendency toward cultural drift when processing non-literal language, often favoring standardized outputs that diverge from the original pragmatic intent. Although Large Language Models (LLMs) have introduced more sophisticated context-handling capabilities, the transition from literal decoding to effective cultural adaptation remains inconsistent. This study investigates these linguistic detours by evaluating ChatGPT-4o, Gemini 1.5 Pro, and Google Translate using a corpus of 100 Brazilian Portuguese expressions. To ensure contemporary relevance, the expressions were validated through *Corpus Carolina* and categorized into four groups: classical idioms, lexically localized expressions, metaphors, and intensifiers. Translation quality was assessed using the Multidimensional Quality Metrics (MQM) framework, focusing on adequacy, fluency, and cultural adaptation. The analysis reveals that, even when grammatical accuracy is achieved, automated systems frequently overlook the sociocultural weight embedded in the source language. Such semantic shifts pose significant challenges in high-stakes professional communication, where nuanced mediation is essential. The findings underscore the limitations of current AI systems in cultural competence and reinforce the ongoing necessity of human intervention to bridge the gap between algorithmic processing and regional identity.

1 Introduction

The rapid advancement of Artificial Intelligence (AI) has significantly reshaped the landscape of Machine Translation (MT), expanding its role in professional, institutional, and organizational communication. Recent advances in Large Language Models (LLMs), particularly those based on attention mechanisms, have led to substantial improvements in fluency and contextual coherence in automated translation systems (Vaswani et al., 2017). Despite these advances, the translation of non-literal language—particularly idiomatic expressions, sociolinguistically marked expressionss, metaphors, and pragmatic intensifiers—continues to pose substantial challenges. These linguistic elements are deeply embedded in cultural and social contexts, making them especially vulnerable to semantic distortion when processed by automated systems. While contemporary studies emphasize the growing contextual sensitivity of LLM-based translators, a persistent gap remains between grammatical accuracy and effective cultural adaptation. Automated systems often prioritize standardized or neutralized outputs, leading to what can be described as cultural drift: a gradual displacement of the original pragmatic intent and socio-cultural resonance of the source text. Such deviations may appear subtle at the lexical level but can generate significant communicative breakdowns, particularly in high-stakes professional environments where meaning negotiation and cultural awareness are essential. This issue is especially relevant in multilingual organizational settings, where translation mediates not only linguistic exchange but also institutional credibility, interpersonal relationships, and decision-making processes. In professional fields such as Executive Secretariat, translation tasks frequently involve culturally marked language used in negotiations, internal communication, and external representation. In these contexts, inadequately adapted translations may compromise clarity, tone, and cultural appropriateness, reinforcing the need for critical human mediation even in technologically advanced translation workflows. Against this backdrop, the present study investigates the limitations of automated cultural adaptation in MT by conducting a comparative evaluation of three widely used systems: ChatGPT-4o, Gemini 1.5 Pro, and Google Translate. The analysis is based on a corpus of 100 Brazilian Portuguese expressions, validated

Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)
Vol. 2, pages 220–227

through the *Corpus Carolina* (Projeto Corpus Carolina, 2020) to ensure contemporary and authentic language use. The expressions are categorized into four linguistic groups—classical idioms, lexically localized expressions, metaphors, and intensifiers—allowing for a systematic examination of how different forms of non-literal language are handled by each system. Translation quality is assessed using the Multidimensional Quality Metrics (MQM) framework, with particular emphasis on adequacy, fluency, and cultural adaptation. By identifying recurrent patterns of semantic shift and cultural neutralization across the evaluated systems, this study aims to delineate the boundaries of current AI-driven translation technologies. The findings contribute to ongoing discussions on the cultural competence of MT systems and underscore the enduring importance of human expertise in ensuring communicative effectiveness across languages and cultures.

2 Methodology

This study adopts an applied, descriptive, and comparative research design aimed at evaluating the capacity of automated translation systems to handle culturally and sociolinguistically embedded, non-literal language. The analysis focuses on identifying patterns of semantic shift and cultural neutralization produced by different Machine Translation (MT) architectures when translating Brazilian Portuguese expressions into English.

2.1 Corpus Selection and Validation

The dataset consists of 100 non-literal expressions originally formulated in Brazilian Portuguese. These expressions were selected to represent recurrent forms of culturally and socially embedded language commonly used in both professional and informal communication contexts. To ensure linguistic authenticity and contemporary relevance, all expressions were validated through the *Corpus Carolina*, a large-scale corpus of Brazilian Portuguese that reflects current language usage across multiple registers. This validation step was employed to confirm that the selected expressions are attested in real communicative contexts rather than artificially constructed examples. Following validation, the expressions were organized into four analytically grounded linguistic categories:

(i) Idiomatic expressions, defined as multiword units whose meanings are not compositionally derived from their individual lexical components (Tagnin, 2013);

(ii) Sociolinguistically marked expressions, referring to expressions associated with specific social groups, communicative contexts, or identities, reflecting structured linguistic variation rather than deviation from a standard norm (Bagno, 2007);

(iii) Metaphorical expressions, understood as linguistic realizations of conceptual mappings between domains, in line with Conceptual Metaphor Theory (Lakoff and Johnson, 1980);

(iv) Intensifiers, defined as lexical or phrasal elements that amplify or modulate the degree of a predicate.

Although some overlap between categories is possible, particularly between idiomatic and sociolinguistically marked expressions, each item was classified according to its dominant linguistic function to ensure analytical consistency. This categorization adopts a descriptive sociolinguistic perspective, in which linguistic variation is treated as a legitimate and meaningful component of language use, rather than as deviation from a standardized variety.

2.2 Evaluated Translation Systems

Three widely used MT systems were selected for comparative evaluation: ChatGPT-4o, Gemini 1.5 Pro, and Google Translate. ChatGPT-4o and Gemini 1.5 Pro represent Large Language Model (LLM) based systems, characterized by their ability to generate context-aware and fluent output through neural generative mechanisms. Google Translate was included as a representative of a conventional MT engine, serving as a baseline for comparison. All translations were generated under default system settings, without user intervention or post-editing, in order to reflect typical real-world usage scenarios. Each source expression was translated independently into English by all three systems.

2.3 Machine Translation Systems and Prompting Strategy

To ensure methodological consistency and avoid bias across translation systems, a minimal and uniform interaction strategy was adopted. For Large Language Models (ChatGPT and Gemini), translations were generated using a direct and uncontextualized prompt equivalent to the instruction "Translate this sentence," followed by the source expression. No additional contextual information, examples, or clarifications were provided. This de-

cision was motivated by the need to maintain comparability with Google Translate, which does not allow prompt-based interaction beyond direct input. By restricting LLM usage to a basic translation command, the study sought to minimize potential advantages associated with prompt engineering and to ensure that observed differences in output could be attributed to the underlying system architectures rather than to interaction design. All translations were generated under default system settings, without manual intervention or post-editing. This approach aligns with the study's objective of evaluating how automated translation systems handle culturally marked expressions under typical user conditions.

2.4 Evaluation Framework

Translation quality was assessed using the Multidimensional Quality Metrics (MQM) framework (Lommel et al., 2014), which enables fine-grained evaluation across multiple linguistic dimensions. Building on the methodological procedures established in the original undergraduate thesis, the present study operationalized MQM through a structured scoring scheme designed to capture both linguistic accuracy and cultural adequacy. For analytical purposes, the evaluation focused on four dimensions: adequacy, fluency, coherence and cohesion, and cultural adaptation. Adequacy refers to the extent to which the translated output preserves the semantic content of the source expression. Fluency assesses grammatical correctness and naturalness in the target language. Coherence and cohesion examine internal textual consistency, while cultural adaptation evaluates whether the translation conveys the pragmatic intent and sociocultural meaning embedded in the original expression. Each dimension was scored on a scale ranging from 0 to 3, where 0 indicates a complete failure to meet the criterion and 3 represents optimal performance. After the evaluation stage, all scores were systematically organized into comparative spreadsheets, allowing both quantitative aggregation and qualitative inspection of recurrent error patterns across categories and systems. The final score of each translated expression was calculated using a simple arithmetic mean, as shown in Equation (1):

$$M = \frac{A + F + C + AC}{4} \tag{1}$$

where M represents the final mean score of the translation, A denotes adequacy, F fluency, C coherence and cohesion, and AC cultural adaptation.

To obtain the overall performance score of each machine translation system (ChatGPT, Gemini, and Google Translate), the individual mean scores of all evaluated expressions were summed and divided by the total number of expressions analyzed, as expressed in Equation (2):

$$M_G = \frac{\sum M_i}{n} \tag{2}$$

where M_G corresponds to the global mean score of the system, M_i refers to the mean score of each individual translation, n represents the total number of evaluated expressions, and $\sum$ indicates summation. This quantitative procedure enabled the identification of numerical performance trends across systems and categories, which were subsequently interpreted through qualitative analysis in the Results and Discussion section, with particular attention to patterns of cultural drift and pragmatic loss.

2.5 Analytical Procedure

The analysis was conducted by comparing source expressions and their respective translations into English within each category. Particular attention was given to cases in which grammatical accuracy was achieved at the expense of cultural or pragmatic equivalence. Examples illustrating successful and unsuccessful adaptations in both target languages were selected to support the discussion of system behavior. This approach enables a nuanced understanding of how different MT architectures handle culturally marked language and highlights the boundaries of automated cultural competence in multilingual professional communication contexts.

2.6 Related Work

Research on Machine Translation (MT) has evolved significantly over recent decades, transitioning from rule-based systems to statistical and neural approaches. While early MT models relied on explicit linguistic rules and bilingual lexicons, limiting their capacity to address ambiguity and contextual variation (Hutchins, 2001), statistical models introduced corpus-driven learning, improving scalability but still facing challenges related to fluency and long-distance dependencies (Koehn, 2010). The emergence of Neural Machine Translation (NMT), particularly through encoder–decoder architectures and attention mechanisms, marked a turning point in the field by

enabling more context-sensitive translations (Bahdanau et al., 2015). The Transformer architecture further advanced this paradigm by relying on self-attention to model linguistic relationships more effectively across languages (Vaswani et al., 2017). More recently, Large Language Models (LLMs) have incorporated translation as part of broader generative frameworks trained on massive multilingual datasets. Although these systems often demonstrate substantial improvements in grammaticality and surface-level coherence, growing evidence suggests that such gains do not necessarily extend to pragmatic adequacy or cultural fidelity. Neural and LLM-based systems have been shown to normalize culturally and sociolinguistically marked expressions, favoring standardized renderings that may obscure pragmatic intent and cultural specificity (Toral et al., 2018). From a sociolinguistic perspective, such normalization may contribute to the attenuation of linguistic diversity and the underrepresentation of socially situated language use. These limitations are particularly salient in the translation of idiomatic expressions, metaphors, and sociolinguistically marked expressions, which rely on shared cultural knowledge rather than compositional meaning. Idiomatic expressions, in particular, have been widely studied in phraseology and are characterized by their non-compositional meaning, as discussed by Tagnin (2013). Within Translation Studies, meaning has long been understood as inseparable from its social and cultural context. From this perspective, translation quality cannot be evaluated solely through linguistic accuracy, but must also consider pragmatic intent, cultural resonance, and communicative purpose (Nida, 1964; House, 2015). Automated translation systems, however, tend to operationalize quality through formal equivalence and fluency-based metrics, often overlooking context-dependent and culturally embedded meanings (Pym, 2010). From a sociolinguistic perspective, linguistic variation is understood as a structured and meaningful phenomenon rather than deviation from a standard norm. In the Brazilian context, studies by Bagno (2007) and Bortoni-Ricardo (2004) emphasize that language reflects social identity, communicative practices, and power relations. This perspective is particularly relevant for the analysis of non-literal and socially marked expressions, which rely on shared cultural knowledge and contextual interpretation. Recent Brazilian scholarship has contributed important empirical and theoretical insights into the use, perception, and evaluation of machine translation in real communicative contexts. Nouatin and Parreiras (Nouatin and Parreiras, 2021) investigate the role of machine translation in the teaching and learning of non-native languages, highlighting teachers' perceptions and attitudes toward automated translation tools. Their findings emphasize that, while MT systems are increasingly present in educational and professional environments, their outputs frequently require critical mediation due to limitations in pragmatic and contextual adequacy. Similarly, Esqueda (Esqueda, 2021) discusses pedagogical and cognitive challenges associated with the use of machine translation, underscoring the need for human interpretive competence to contextualize and refine automated outputs. These Brazilian studies reinforce broader concerns regarding the gap between linguistic fluency and communicative effectiveness in automated translation systems. They also foreground the importance of qualitative evaluation approaches that account for discourse, pragmatics, and cultural meaning, dimensions that are often underrepresented in automatic metrics. In this regard, the Multidimensional Quality Metrics (MQM) framework has emerged as a robust alternative, offering a fine-grained taxonomy of error categories across linguistic, semantic, and pragmatic dimensions (Lommel et al., 2014). MQM enables systematic qualitative analysis of translation behavior, making it particularly suitable for studies concerned with culturally marked language. Corpus-based approaches have likewise played a central role in advancing MT research and evaluation. The use of curated and linguistically validated corpora allows researchers to ground their analyses in authentic language use and contemporary discourse patterns. In the Brazilian context, the *Corpus Carolina* has been employed as a reference for contemporary Brazilian Portuguese usage, providing a reliable basis for the selection and validation of idiomatic and culturally situated expressions (Projeto Corpus Carolina, 2020). Such resources are especially relevant for studies examining non-literal language, as they help ensure that analyzed expressions reflect current linguistic practices rather than prescriptive or outdated forms. Despite growing interest in culturally informed evaluation and the increasing availability of neural and LLM-based translation systems, comparative studies that systematically examine how different MT architectures handle non-literal language across multiple categories remain limited. Existing re-

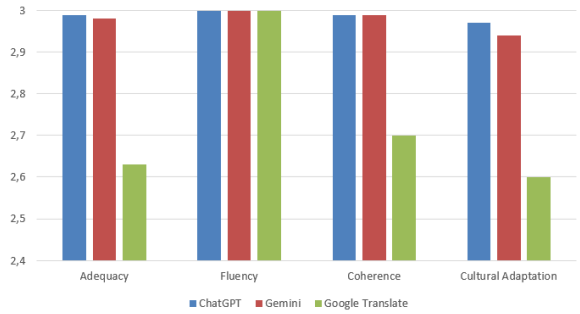


Figure 1: Average dimension scores per tool (Scale 1-3) across the 100-item corpus.

search often prioritizes sentence-level accuracy or post-editing effort, leaving open questions regarding broader patterns of cultural drift in automated translation. This study addresses this gap by analyzing how distinct MT systems process idioms, sociolinguistically marked expressions, metaphors, and intensifiers in Brazilian Portuguese, combining a corpus-based methodology with a qualitatively grounded evaluation framework.

3 RESULTS AND DISCUSSION

3.1 Overall Patterns Observed

This section presents and discusses the results obtained from the MQM-based evaluation of automated translations of Brazilian Portuguese expressions into English. To ensure analytical focus and conciseness, the analysis concentrates on a selected subset of the corpus, consisting of representatives expressions per category (idioms, lexically localized expressions, metaphors, and intensifiers/comparative expressions). While all selected expressions were evaluated according to the same criteria, the presentation of results prioritizes the most illustrative cases in each category, as is common practice in full-length conference papers to avoid excessive extension.

Figure 1 2 presents the mean scores for cultural adaptation obtained by each translation system across all evaluated expressions. The results indicate systematic differences in how the systems handle culturally embedded language. LLM-based systems (ChatGPT-4o and Gemini 1.5 Pro) achieved higher average scores in fluency and adequacy, reflecting their ability to generate grammatically natural and contextually coherent output. However, gains in linguistic quality did not consistently translate into effective cultural adaptation. Google Translate, while exhibiting lower fluency scores overall, demonstrated more stable behavior in literal equivalence, albeit frequently failing to convey pragmatic intent in non-literal expressions. These findings suggest that improvements in surface-level coherence do not necessarily correspond to better handling of cultural meaning, reinforcing concerns raised in recent MT research regarding semantic normalization in neural systems.

3.2 Idiomatic Expressions

Idiomatic expressions constitute one of the most challenging categories for automated translation, as their meanings are not compositionally derived from individual lexical items. Table 1 presents representative examples of idiomatic expressions and their corresponding MQM-based scores.

Expression	System	Translation	Score
Chutar o balde	ChatGPT	give up completely	2.25
	Gemini	lose patience	2.00
	Google Translate	kick the bucket	0.75
Acabar em pizza	ChatGPT	end with no consequences	2.50
	Gemini	end in nothing	1.75
	Google Translate	end in pizza	0.50
Engolir sapo	ChatGPT	put up with something unfair	2.25
	Gemini	swallow an insult	2.00
	Google Translate	swallow a frog	0.75

Table 1: MQM scores for selected idiomatic expressions

Table 2 summarizes the translation results for the selected idiomatic expressions (*Chutar o balde, Acabar em pizza, Engolir sapo, Amigo da onça, Quebrar o galho*). Across systems, idioms posed substantial challenges, particularly with respect to cultural adaptation. Literal translations were frequent, especially in cases where idiomatic equivalents exist in the target language but require pragmatic inference rather than compositional decoding. ChatGPT-4o demonstrated relatively higher adequacy scores for idioms with close functional equivalents, such as *Quebrar o galho*. Nevertheless, even in these cases, translations often favored explanatory paraphrases over idiomatic substitutions, reducing pragmatic force. Gemini 1.5 Pro exhibited similar tendencies, while Google Translate consistently produced literal renderings, resulting in lower cultural adaptation scores. These results align with previous findings that idiomaticity remains a persistent weakness in automated translation systems.

3.3 Sociolinguistically Marked Expressions

Sociolinguistically marked expressions reflect patterns of linguistic variation associated with specific social groups, communicative contexts, and cultural practices. From a sociolinguistic perspective, such variation is understood as a structured and meaningful component of language use rather than deviation from a standardized norm (Bagno, 2007; Bortoni-Ricardo, 2004). These expressions are therefore especially sensitive to normalization strategies in automated translation. Table 2 summarizes representative translations of Brazilian Portuguese expressions exhibiting sociolinguistic marking. In this study, these expressions are identified based on their distribution across communicative contexts and their attestation in corpus data, rather than on prescriptive or geographically restrictive criteria. In this study, the term “regionalism” and "sociolinguistically marked expressions" is not employed in a normative or hierarchical sense; instead, it functions as an operational label for lexical items with predominantly localized distribution, supported by corpus-based evidence.

Expression	System	Translation	MQM Mean
Oxente	ChatGPT	wow/really?	2.00
	Gemini	what?	1.75
	Google Translate	oxente	0.50
Arretado	ChatGPT	impressive	2.25
	Gemini	intense	1.75
	Google Translate	angry	1.00
Migué	ChatGPT	an excuse/deception	2.00
	Gemini	trick	1.75
	Google Translate	migué	0.50

Table 2: MQM mean scores for selected Brazilian Portuguese regionalisms.

The analysis reveals a strong tendency toward cultural neutralization. While LLM-based systems attempted contextual approximation, they frequently diluted regional identity. Google Translate systematically failed to interpret sociolinguistically marked expressions, often leaving terms untranslated or assigning semantically unrelated meanings.

3.4 Metaphorical Expressions

Metaphorical expressions displayed intermediate levels of difficulty. As shown in Table 3, some metaphors were rendered through literal transfer, while others were paraphrased, affecting expressive intensity.

Expression	System	Translation	MQM Mean
Luz no fim do túnel	ChatGPT	light at the end of the tunnel	2.75
	Gemini	hope ahead	2.25
	Google Translate	light at the end of the tunnel	2.50
Fogo de palha	ChatGPT	short-lived enthusiasm	2.25
	Gemini	something temporary	2.00
	Google Translate	straw fire	1.00

Table 3: MQM mean scores for selected metaphorical expressions.

The final category revealed marked divergence across systems. Highly figurative comparisons posed particular difficulties, frequently resulting in either literal translations or neutral paraphrases. ChatGPT-4o occasionally generates culturally adapted equivalents, especially for expressions with approximate analogs in English. However, these adaptations were inconsistent and often depended on implicit inference rather than systematic cultural mapping. Google Translate, by contrast, largely failed to capture the evaluative force of intensifiers, producing translations that conveyed factual meaning without pragmatic emphasis.

3.5 Cross-System Comparison and Implications

Overall, ChatGPT-4o and Gemini demonstrated greater contextual flexibility, achieving higher adequacy and fluency scores. However, both systems exhibited a tendency toward cultural neutralization, which reduced region-specific and idiomatic features. Google Translate, while more consistent in literal transfer, showed limited capacity for pragmatic interpretation. These patterns highlight the limitations of current automated translation systems in high-stakes communicative contexts, where cultural mediation plays a central role. The findings reinforce the need for human intervention in

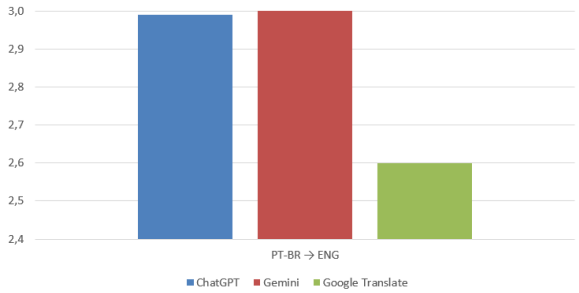


Figure 2: Mean Cultural Adaptation scores per MT system

professional domains, where accurate interpretation of culturally embedded language is essential for effective communication.

3.6 Examples of Cultural Mismatch

MT tools often fail to preserve the cultural nuances of source texts, leading to misinterpretations or loss of intended meaning. Table 2 illustrates some examples of Brazilian Portuguese expressions translated into English, highlighting cases of cultural mismatch and inadequate adaptation. These examples demonstrate how literal translations or lack of context can distort the original message.

Original Expression	Tool	Translation	Justification
Moscou	ChatGPT	Moscow	Confuses Brazilian slang ("distraiu / vacilou") with proper noun "Moscow." Ideal: "He/She zoned out / slipped up."
Chutar o balde	ChatGPT	Kick the bucket	Literal translation misleads; English idiom refers to dying instead of "giving up."
Ficar de molho	Google Translate	Stay in sauce	Literal translation loses the meaning of resting or recovering.
Pagar o pato	ChatGPT	Pay the duck	Literal translation loses idiomatic meaning ("take the blame").
Dar com os burros n'água	Gemini	Give the donkeys in the water	Literal translation misinterprets the expression meaning ("fail").

Table 4: Examples of cultural mismatch in machine translation

3.7 Professional and Organizational Implications

Although artificial intelligence models demonstrate superior performance in terms of fluency and processing speed, the findings of this study indicate that they do not fully replace human judgment in tasks involving culturally nuanced language. It is important to emphasize that this outcome is directly related to the methodological choices adopted, as translations were generated from isolated expressions, without sentential context, and through minimal prompts, with no iterative refinement or interaction. These controlled methodological decisions highlight how the absence of linguistic and pragmatic context can significantly affect the cultural interpretation of non-literal expressions by automated systems. These results support the argument proposed by Amini et al. (2024), who frame machine translation as a collaborative rather than a substitutive technology. In this perspective, automated systems function as productivity-enhancing tools whose outputs still require human mediation, particularly in communicative situations where pragmatic intent and cultural meaning are central. The implications of these findings are especially relevant for professionals engaged in multilingual information management and intercultural communication, including translators, interpreters, language educators, and, notably, Executive Secretariat professionals. As emphasized by Florido (2021), executive secretaries act as cultural mediators and strategic information managers within organizational contexts, playing a key role in companies that operate internationally. Consequently, developing critical awareness of the limitations and affordances of machine translation systems becomes an essential professional competence for ensuring communicative accuracy and cultural adequacy in documents, meetings, negotiations, and corporate correspondence. From an organizational perspective, the results suggest that companies operating in multicultural environments may benefit from adopting structured evaluation frameworks such as MQM to assess the quality of machine-translated content used in reports, emails, contracts, and institutional materials. Such practices contribute to the standardization of intercultural communication and reduce risks associated with misinterpretation. Furthermore, the strategic use of well-designed prompts in AI-based translation tools can improve semantic and pragmatic alignment with the intended com-

municative context. Nevertheless, even when productivity gains and cost reductions are achieved, effective deployment of these technologies still depends on continuous monitoring and qualified human mediation to ensure that culturally embedded meanings are preserved throughout the translation process.

4 Conclusion

This study investigated how contemporary automated translation systems handle culturally embedded Brazilian Portuguese expressions, focusing on idioms, sociolinguistically marked expressions, metaphors, and intensifiers. Using a corpus-based selection validated through the *Corpus Carolina* and an MQM-inspired evaluation framework, the analysis revealed that advances in fluency and grammatical accuracy do not necessarily translate into effective cultural adaptation. Even when translations were semantically adequate, pragmatic intent and socio-cultural resonance were frequently attenuated. The results demonstrate that LLM-based systems such as ChatGPT-4o and Gemini 1.5 Pro generally outperform conventional MT engines in terms of fluency and overall adequacy. However, these systems often favor paraphrasing and semantic normalization, which reduces idiomatic force and cultural specificity. Google Translate, while more consistent in literal equivalence, systematically failed to convey non-literal meaning, particularly in expressions that depend on shared cultural knowledge. From a professional perspective, especially in fields that require intercultural mediation, these findings highlight the continued necessity of human intervention in high-stakes communication. Automated translation tools function more effectively as support technologies rather than substitutes for culturally informed mediation. Future research may expand the corpus size, explore additional language pairs, and investigate prompt-sensitive evaluation strategies to further examine the boundaries of automated cultural adaptation.

References

Marcos Bagno. 2007. *Preconceito linguístico: o que é, como se faz*. Loyola.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of the 3rd International Conference on Learning Representations*.

Stella Maris Bortoni-Ricardo. 2004. *Educação em língua materna: a sociolinguística na sala de aula*. Parábola.

Marileide Dias Esqueda. 2021. Machine translation: teaching and learning issues. *Trabalhos em Linguística Aplicada*, 60(1):282–299.

Juliane House. 2015. *Translation Quality Assessment: Past and Present*. Routledge.

John Hutchins. 2001. Machine translation and human translation: In competition or cooperation? *International Journal of Translation*, 13(1).

Philipp Koehn. 2010. *Statistical Machine Translation*. Cambridge University Press.

George Lakoff and Mark Johnson. 1980. *Metaphors We Live By*. University of Chicago Press.

Arle Lommel, Aljoscha Burchardt, and Hans Uszkoreit. 2014. Multidimensional quality metrics (mqm): A framework for declaring and describing translation quality metrics. In *Proceedings of the 9th International Conference on Language Resources and Evaluation*.

Eugene A. Nida. 1964. *Toward a Science of Translating*. Brill.

Gbènoukpo Gérard Nouatin and Vicente Aguimar Parreiras. 2021. Tradução automática no ensino e na aprendizagem de línguas não maternas: percepções, atitudes e opiniões de professores. *Trabalhos em Linguística Aplicada*, 60(3):841–852.

Projeto Corpus Carolina. 2020. Corpus carolina: um corpus de referência do português brasileiro contemporâneo. Disponível para pesquisa linguística e validação de uso contemporâneo.

Anthony Pym. 2010. *Exploring Translation Theories*. Routledge.

Stella Esther Ortweiler Tagnin. 2013. *O jeito que a gente diz: combinações consagradas em inglês e português*. Disal.

Antonio Toral, Sheila Castilho, Ke Hu, and Andy Way. 2018. Post-editing effort of neural machine translation: A study of human factors. *Machine Translation*, 32(1–2).

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*.

Towards a Universal Dependencies Corpus for Portuguese Epidemiological Reports

Christian Freitas[1], Livy Real[2,3], Lilian Berton[1], Valeria de Paiva[4]

[1]Universidade Federal de São Paulo, São Paulo, Brazil
[2]Universidade Federal do Amazonas, Manaus, Brazil
[3]Instituto Kunumi, Belo Horizonte, Brazil
[4]Topos Institute, Berkeley, USA

christian.freitas@unifesp.br · livy@kunumi.com · lberton@unifesp.br · valeria@topos.institute

Abstract

We present an ongoing research project focused on the construction of a Universal Dependencies (UD) corpus of Portuguese epidemiological reports derived from documents published within the Brazilian public health system. We describe findings and challenges to build such a corpus from PDF reports processed through a controlled document extraction pipeline that contrasts layout-aware extraction with raw PDF text extraction, explicitly addressing the impact of tabular content on downstream syntactic analysis. Narrative text is annotated using multiple UD parsers for Portuguese, including widely used and state-of-the-art tools, and their outputs are systematically compared using descriptive structural indicators and targeted qualitative inspection.

Our analysis highlights domain-specific challenges in epidemiological texts and shows that document extraction and representation choices have a stronger effect on parsing behavior than parser selection alone. Based on these findings, we identify robust preprocessing configurations and discuss design choices for a UD-epidemiological corpus to support future research on syntactic parsing, domain adaptation, and downstream natural language processing tasks in epidemiology and public health.

1 Introduction

Universal Dependencies (UD) has become a widely adopted framework for syntactic annotation, enabling cross-linguistic consistency and facilitating the development and evaluation of dependency parsers across languages and domains (Nivre et al., 2016, 2020). For Portuguese, several UD treebanks and parsing tools have been developed, supporting a wide range of natural language processing applications (Rademaker et al., 2017; Branco et al., 2022; Sanches Duran et al., 2025). Despite these advances, most existing resources and evaluations focus on newswire or general-domain texts, leaving specialized domains comparatively underexplored. Exceptions include Di Felippo et al. (2024); Souza and Freitas (2023).

Epidemiological reports constitute a particularly challenging domain for syntactic parsing. Such documents typically combine technical terminology, numerical expressions, abbreviated forms, and complex syntactic constructions, often embedded in heterogeneous document layouts that include tables, lists, and scanned pages. These characteristics can negatively impact both text extraction and downstream syntactic analysis, especially when models trained on general-domain data are applied without adaptation.

We intend to address this gap by producing a Universal Dependencies corpus of Portuguese epidemiological reports derived from documents published within the Brazilian public health surveillance system. In this work, we focus on a preliminary step for the construction of the corpus through a controlled document processing pipeline that compares different text extraction strategies, evaluates the impact of optical character recognition when required, and applies multiple UD parsers for Portuguese, including state-of-the-art and widely used tools. Our goal is not to propose new parsing models, but rather to quantify parsing behavior and common error patterns in this specialized domain and to release a curated corpus that can support future research on syntactic analysis and domain adaptation. Our long-term goal is to produce a reliable UD corpus for epidemiological reports in Portuguese.

2 Epidemiological Reports

The SIREVA (Sistema Regional de Vacinas) system is a public health surveillance initiative coordinated in Brazil within the Unified Health System in Portuguese, Sistema Único de Saúde (SUS). The SIREVA-SUS system focuses on the monitoring of invasive bacterial diseases and vaccine-preventable

Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)
Vol. 2, pages 228–237

pathogens. The system produces periodic epidemiological reports that consolidate laboratory-confirmed cases, serotype distributions, and temporal and geographic trends, serving as an important source of information for epidemiological understanding and public health decision-making.

The documents analyzed in this work consist of official SIREVA-SUS epidemiological reports, which are published in PDF format. These reports typically combine continuous narrative text with tables, lists, and summary statistics. They may include scanned pages depending on the publication year and source. From a natural language processing perspective, this heterogeneous structure poses challenges for automatic text extraction and syntactic analysis, as layout artifacts and domain-specific formatting can negatively affect tokenization, sentence segmentation, and dependency parsing. Figure 1 illustrates this contrast using content extracted directly from the 2024 SIREVA-SUS report: a representative structured data table alongside its corresponding narrative interpretation.

Tabela 1. Número de isolados invasivos por grupo etário e sexo

Grupo etário	*Sexo*							
	Masculino		Feminino		Sem dado		Total	
	n	%	n	%	n	%	n	%
< 12 meses	25	52,1	21	43,8	2	4,2	48	18,1
12–23 meses	11	57,9	8	42,1	0	0,0	19	7,2
24–59 meses	17	56,7	13	43,3	0	0,0	30	11,3
Subtotal (1)	**53**	**54,6**	**42**	**43,3**	**2**	**2,1**	**97**	**36,6**
5–14 anos	13	46,4	15	53,6	0	0,0	28	10,6
15–29 anos	14	66,7	7	33,3	0	0,0	21	7,9
30–49 anos	18	56,3	14	43,8	0	0,0	32	12,1
Subtotal (2)	**45**	**55,6**	**36**	**44,4**	**0**	**0,0**	**81**	**30,6**
50–59 anos	10	40,0	15	60,0	0	0,0	25	9,4
≥ 60 anos	26	41,9	36	58,1	0	0,0	62	23,4
Subtotal (3)	**36**	**41,4**	**51**	**58,6**	**0**	**0,0**	**87**	**32,8**
Total	**134**	**50,6**	**129**	**48,7**	**2**	**0,8**	**265**	**100,0**

*"Do total de **265** amostras H. influenzae: **166** se referem à cultura e **99** se referem a PCR em tempo real."*

Figure 1: Example of content from the 2024 SIREVA-SUS report: a structured data table (top) followed by its narrative interpretation (bottom). This juxtaposition illustrates the heterogeneous nature of epidemiological documents, where tabular evidence and textual claims coexist and must be handled separately by the processing pipeline.

In this study, the SIREVA-SUS reports serve as a representative example of real-world epidemiological documents in Portuguese. By focusing on this data source, we aim to evaluate the behavior of UD parsers under domain-specific conditions and to construct a corpus that reflects the linguistic and structural characteristics commonly found in epidemiological surveillance reports.

Table 1: Pages and heuristically detected tables per SIREVA-SUS report.

Year	**Pages**	**Tables**
2013	42	51
2014	41	53
2015	43	51
2016	43	89
2017	41	62
2018	41	61
2019	38	61
2020	36	52
2021	37	53
2022	37	58
2023	43	69
2024	42	51
Total	**484**	**711**

The documents analyzed in this work consist of official SIREVA-SUS epidemiological reports, which are published in PDF format and made publicly available by the Adolfo Lutz Institute.[1]

3 Methodology

Before constructing the full corpus, we adopt a pilot-based evaluation strategy in which the entire document processing and parsing pipeline is applied to a single, representative epidemiological report. Specifically, all experiments reported in this section are conducted using the SIREVA-SUS report from 2024.

This document was selected because it is structurally representative of the collection as a whole, combining narrative text, extensive tabular content, and modern PDF formatting. By focusing initially on a single report, we are able to analyze the effects of document extraction choices, table handling, and parser behavior in a controlled setting, while avoiding confounding variation introduced by inter-document heterogeneity.

We therefore adopt a representation strategy that separates text and tables *physically*, while preserving their *logical connections* at the document level. Conceptually, each report is modeled as a collection of textual statements and tabular evidence objects, linked by explicit relations that capture their rhetorical and evidential roles.

[1] https://www.ial.sp.gov.br/ial/publicacoes/boletim

3.1 Text and Table Extraction

During preprocessing, narrative text blocks (e.g., paragraphs, section summaries) and tables are extracted independently from the original PDF documents. Text blocks are segmented into coherent units (typically paragraphs) and assigned stable identifiers. Tables are parsed into structured representations that preserve row and column headers, cell values, units, captions, and footnotes. Numeric values are retained in their original form and are not subjected to language modeling.

This separation ensures that tables remain amenable to deterministic processing, validation, and normalization, while text blocks remain suitable for natural language processing techniques.

3.2 Tables as Structured Evidence

Each table is treated as a structured evidence object rather than as a textual artifact. Rows and columns are interpreted according to their semantic roles (e.g., disease, year, geographic region, measure type), and individual cell entries correspond to atomic factual statements. Captions and table-local annotations are preserved as metadata, as they often specify measurement conventions, exclusions, or temporal scope.

By maintaining tables in a structured form, the pipeline supports downstream tasks such as unit normalization, consistency checking, aggregation, and cross-document comparison, which are essential in medical and epidemiological settings.

3.3 Linking Textual Claims and Tables

To capture the intended relationship between narrative and data, explicit links are introduced between text blocks and tables.These links are inferred automatically using a combination of explicit textual references (e.g., "Table 3 shows..."), document structure, proximity heuristics, and caption semantics. Each link is labeled with one of three coarse-grained relation types: *supported_by* (the narrative claim is directly backed by tabular data), *elaborates* (the text expands on information presented in a table), and *summarizes* (the text condenses tabular content into a higher-level statement). Manual validation of these automatically inferred links is planned as part of the corpus curation process described in Section 7.

This yields a document-level representation in which textual claims and tabular evidence are connected but remain distinct. Such a representation makes it possible to trace which quantitative data support which assertions, to identify uncited or weakly supported claims, and to reason jointly over text and data without conflating their roles.

3.4 Implications for Information Extraction

Under this representation, information extraction from tables is performed deterministically over structured data, while UD processing is applied primarily to narrative text, captions, and annotations. Large language models, when used, are restricted to interpreting metadata and mapping textual descriptions to canonical schemas, and are never treated as the source of numeric ground truth.

This separation-of-concerns design reduces error propagation, supports validation and provenance tracking, and reflects the epistemic distinction between quantitative evidence and its narrative interpretation. In oversight and reporting contexts, this distinction is critical for ensuring transparency and analytical reliability.

4 Parsing Configurations and Tool Selection

Why Universal Dependencies. We adopt the Universal Dependencies (UD) framework as the syntactic representation for this corpus for three reasons: cross-resource comparability, parser diversity, and structural transparency. First, UD provides a linguistically motivated but application-agnostic annotation scheme that is consistent across languages and domains, allowing the resulting corpus to be compared directly with existing Portuguese treebanks and with epidemiological corpora in other languages. Second, most widely used dependency parsers for Portuguese either natively produce UD annotations or can be reliably mapped to UD, making it possible to evaluate multiple parsing configurations within a shared representational space. Finally, UD's explicit encoding of predicate–argument structure, modification, coordination, and clause boundaries makes it well suited for analyzing the syntactic realizations of epidemiological statements, such as causal claims, temporal descriptions, and quantified assertions. Our goal is not to advance syntactic theory, but to obtain a stable, interpretable syntactic layer that supports error analysis and future downstream tasks, including information extraction and argument-level modeling.

Why CoNLL-U. All parsed outputs are normalized to the CoNLL-U format, the official in-

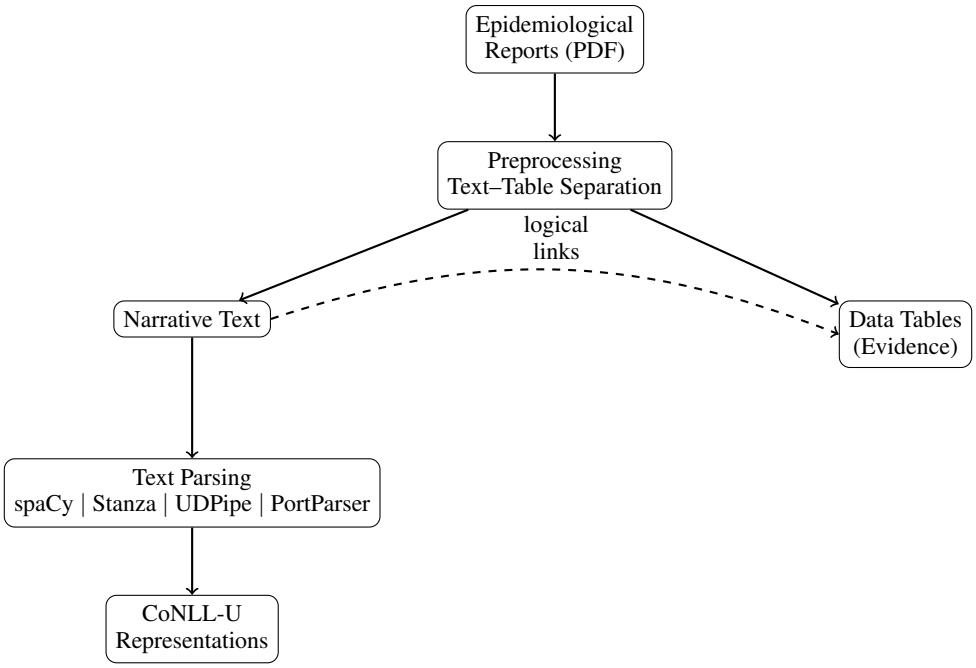


Figure 2: Processing pipeline for epidemiological reports. Reports are preprocessed to separate narrative text from tabular data while preserving their logical connections. Narrative text is parsed using multiple dependency parsers and converted to CoNLL-U format, while tables are retained as structured evidence objects and excluded from linguistic parsing.

terchange format of the Universal Dependencies project. CoNLL-U provides a compact but expressive representation that integrates tokenization, lemmatization, morphological features, part-of-speech tags, and dependency relations in a single, line-oriented structure. Using CoNLL-U enables direct comparison between parsers with different internal architectures and preprocessing strategies, while preserving sufficient linguistic detail for qualitative inspection and downstream reuse. In addition, CoNLL-U supports partial annotation, sentence-level alignment, and post-hoc correction, which is essential when working with automatically extracted text from noisy PDF sources. By committing to CoNLL-U, we ensure that the released corpus can be easily inspected, extended, or re-parsed as improved models or domain-adapted parsers become available.

5 Document Extraction with Docling

Docling (IBM Research, 2024) is a document processing framework designed to extract structured textual content from complex PDF documents, with particular emphasis on preserving layout information and separating different content types, such as continuous text and tables. Unlike raw PDF text extraction methods, which operate primarily at the character or line level, Docling performs layout-aware segmentation, allowing for a more faithful reconstruction of document structure.

In the context of epidemiological reports, which frequently combine narrative descriptions with tabular data and heterogeneous formatting, layout-aware extraction is especially relevant. By explicitly distinguishing between running text and tables, Docling enables downstream linguistic processing to focus on syntactically meaningful textual units, while avoiding artifacts commonly introduced by table structures when treated as plain text.

In this work, Docling is used as the primary document extraction method and is compared against a raw PDF text extraction baseline. This comparison allows us to assess the impact of layout-aware document processing on subsequent UD parsing quality and to quantify the extent to which document structure influences syntactic analysis in epidemiological texts.

5.1 Removal of Markdown Structural Punctuation in Docling Output

During initial experiments, we observed that Docling exports textual content from PDF documents using a Markdown-based representation, particularly when encoding tabular structures. In this representation, structural elements such as vertical

bars (|), header markers (#), and column separators become part of the linearized text.

This behavior had a direct impact on downstream analyses. Markdown-specific characters were tokenized by the linguistic parsers (spaCy, Stanza, and UDPipe), leading to an artificial inflation in the number of tokens, lexical types, and sentences. As a consequence, medically relevant terms (e.g., *meningite*, *pneumonia*) were repeatedly counted in contexts that do not correspond to natural language usage, but rather to the structural encoding of tables.

To mitigate this effect and ensure comparability with other PDF text extraction methods (such as PyPDF and PyMuPDF), we introduced an explicit post-processing step for text extracted with Docling. The goal of this step was to preserve the informational content conveyed by tables while eliminating Markdown-specific structural punctuation that does not carry linguistic meaning.

Adopted Strategy. The adopted strategy consists of removing typical Markdown characters used for table and title formatting, including vertical bars, header markers, column separator lines (e.g., `---`), and redundant punctuation introduced by the Markdown layout. Only the textual and numerical content of table cells is retained, in a linearized form comparable to the output produced by traditional PDF text extractors.

Comparative analyses across extraction variants show that the original Docling output with Markdown markup yields substantially higher token and sentence counts. After the removal of Markdown structural punctuation, these values closely align with those observed for other PDF readers. This confirms that the observed explosion in token and sentence counts is not driven by semantic content, but by the structural representation of tables.

These findings indicate that current linguistic parsers are not designed to operate directly on Markdown-encoded tabular structures, reinforcing the need for careful preprocessing when technical documents rich in tables are used as input to syntactic parsers.

6 UD Parsers

Given the choice of Universal Dependencies as the syntactic framework, the selection of parsing tools follows naturally. We focus on parsers that (i) natively produce UD-compliant analyses, (ii) support Portuguese with publicly available models, and (iii) differ substantially in architectural design, training data, and intended use. This allows us to examine how distinct parsing paradigms behave when applied to epidemiological text extracted from complex PDF documents, while holding the annotation scheme and output format fixed. By normalizing all parser outputs to CoNLL-U, we ensure that observed differences reflect genuine parsing behavior rather than representational incompatibilities. The resulting comparison is therefore not a competition between systems, but a controlled evaluation of robustness, error patterns, and domain sensitivity under a shared syntactic standard.

To assess the robustness of dependency parsing in the epidemiological domain, we employ three widely used UD parsers for Portuguese: spaCy, Stanza, and UDPipe. These tools represent different design choices and levels of linguistic modeling, and are commonly adopted as baselines in dependency parsing evaluations. In addition, we include the PortParser (Lopes and Pardo, 2024), a state-of-the-art dependency parser for Portuguese.

spaCy. spaCy is an industrial-strength natural language processing library that provides efficient pipelines for tokenization, part-of-speech tagging, and dependency parsing (Honnibal et al., 2020). Its dependency parser is based on transition-based neural models optimized for speed and scalability. Although spaCy is not primarily designed for linguistic research, it supports Universal Dependencies labels and is frequently used in applied NLP settings. In this work, spaCy serves as a pragmatic baseline, allowing us to evaluate how a general-purpose, high-performance parser behaves when applied to specialized epidemiological texts. spacy offers a library to produce Universal Dependencies in a CNLL format. Note though that spaCy was developed for English and its Portuguese models are not particularly fine-tuned for medical data.

Stanza. Stanza (Qi et al., 2020) is a neural NLP toolkit developed by the Stanford NLP Group, designed with a strong focus on linguistic accuracy and multilingual support. It provides end-to-end pipelines for tokenization, morphological analysis, part-of-speech tagging, lemmatization, and dependency parsing, all trained within the Universal Dependencies framework. Stanza's models rely on deep contextualized representations and have shown competitive performance across multiple languages and treebanks.

UDPipe. UDPipe (Straka et al., 2016) is a trainable pipeline for processing text in the CoNLL-U format, offering models for tokenization, tagging, lemmatization, and dependency parsing. It has been extensively used in shared tasks and benchmark studies related to UD, and is known for its efficiency and reproducibility. UDPipe models are trained directly on UD treebanks and follow the official annotation guidelines closely, making the tool particularly suitable for comparative evaluations. In this study, UDPipe provides a strong and well-established baseline for assessing parsing quality in Portuguese epidemiological reports.

PortParser. PortParser (Lopes and Pardo, 2024) is specifically designed for Portuguese and leverages recent advances in neural dependency parsing, achieving top performance on standard Portuguese UD benchmarks. Its architecture and training strategy are optimized to capture language-specific syntactic phenomena that are often underrepresented in multilingual or general-purpose models.

The inclusion of PortParser allows us to establish a reference for parsing quality in Portuguese and to assess how models behave when applied to a specialized and out-of-domain setting such as epidemiological reports. By comparing PortParser with more general-purpose UD parsers, we aim to identify whether gains observed in benchmark evaluations transfer to real-world epidemiological texts, which exhibit domain-specific terminology, numerical expressions, and heterogeneous document structures.

In this section, we analyze the behavior of all parsers (spaCy, Stanza, UDPipe, and PortParser v2) across different text extraction scenarios using basic structural metrics: number of sentences, number of tokens, number of word-form types, and number of lemma types. Rather than ranking parsers by performance, our objective is to ground the discussion in observed quantitative differences and to understand how parser design choices interact with document preprocessing decisions in the syntactic analysis of epidemiological reports.

Table 2 summarizes the main structural statistics for each combination of extraction scenario and parser, providing the empirical basis for the analyses discussed below.

Text extraction scenarios. Table 2 summarizes four text extraction settings designed to isolate the effects of (i) raw PDF extraction versus layout-aware extraction and (ii) the presence of linearized tabular content in the parser input. We use the following scenario labels throughout the results: **A_raw_pypdf** (raw text extracted with PyPDF), **D_raw_pymupdf** (raw text extracted with PyMuPDF), **B_docling_text_only** (Docling layout-aware extraction that still retains linearized table content), and **B2_docling_text_only_no_tables** (Docling extraction with explicit removal of tables, keeping only running narrative text).

Lexical stability (tokens and types). As shown in Table 2, spaCy, Stanza, and PortParser v2 exhibit highly similar behavior with respect to the total number of tokens, word-form types, and lemma types under raw extraction scenarios (PyPDF and PyMuPDF). In these settings, all three parsers converge to nearly identical values, indicating stable tokenization and lemmatization when the input text does not contain complex tabular structures or artificial markup.

UDPipe, in contrast, displays a markedly different lemmatization profile. Although its token counts and word-form type counts are comparable to those of the other parsers, the number of distinct lemmas is consistently lower across all scenarios. This pattern, visible in both raw and Docling-based extractions, suggests a more aggressive normalization strategy or a less fine-grained lemmatization process, which may impact downstream tasks that depend on lexical diversity.

Sentence segmentation variability. Sentence segmentation shows the largest variation across parsers. For identical input text, UDPipe consistently produces the highest number of sentences, while spaCy yields intermediate values and Stanza produces fewer sentences. PortParser v2 exhibits the most conservative segmentation behavior, generating the smallest number of sentences in most scenarios (Table 2).

These differences directly affect the granularity of linguistic units and have implications for downstream tasks such as information extraction, text–table alignment, discourse analysis, and the modeling of long-range syntactic and semantic relations.

Impact of text extraction and preprocessing. Beyond parser-specific behavior, Table 2 shows that the choice of text extraction method exerts an effect that is comparable to, and in some cases greater than, the choice of parser. In the Docling extraction scenario that retains linearized ta-

Table 2: Descriptive statistics of parsed outputs across document extraction variants and dependency parsers, including PortParser v2.

Scenario	Parser	Sentences	Tokens	Types (form)	Types (lemma)
A_raw_pypdf	PortParser v2	276	12193	1773	1747
A_raw_pypdf	spaCy	724	11347	1762	1732
A_raw_pypdf	Stanza	471	11297	1775	1755
A_raw_pypdf	UDPipe	1273	12405	1781	833
B2_docling_text_only_no_tables	PortParser v2	95	1603	363	346
B2_docling_text_only_no_tables	spaCy	145	1559	370	348
B2_docling_text_only_no_tables	Stanza	172	1575	356	341
B2_docling_text_only_no_tables	UDPipe	163	1607	354	188
B_docling_text_only	PortParser v2	1115	28067	1787	1758
B_docling_text_only	spaCy	2304	21477	1805	1774
B_docling_text_only	Stanza	290	21420	1830	1810
B_docling_text_only	UDPipe	1689	23114	1841	817
D_raw_pymupdf	PortParser v2	276	12186	1771	1745
D_raw_pymupdf	spaCy	648	11334	1759	1729
D_raw_pymupdf	Stanza	480	11284	1773	1753
D_raw_pymupdf	UDPipe	1294	12395	1779	831

bles, all parsers exhibit a substantial increase in token counts and a pronounced inflation in sentence counts. This effect is particularly evident for spaCy and UDPipe, whose sentence counts increase by an order of magnitude relative to cleaner extraction settings.

When tables are explicitly removed, (`docling_text_only_no_tables`), parser behavior returns to patterns closely aligned with those observed under raw PDF extraction. This confirms that the degradation in parsing stability is driven not by the narrative content itself, but by the representational form of tabular data.

This effect arises from the presence of linearized tabular content, which introduces non-natural patterns of punctuation, repetition, and structural markers that interfere with both tokenization and sentence boundary detection. When tables are explicitly removed, parser behavior returns to patterns similar to those observed under raw PDF extraction. This confirms that it is not the narrative text itself that degrades parsing behavior, but rather the representational form of tabular data.

Sentence fragmentation under noisy extraction. Figure 3 complements Table 2 by normalizing sentence counts per 1k tokens. This visualization highlights that extraction pipelines including linearized tables consistently lead to higher sentence fragmentation across all parsers. Although the magnitude of the effect varies by parser, the overall trend is stable: noisy extraction amplifies segmentation artifacts independently of parser architecture.

Morphosyntactic profile and structural noise. Beyond sentence-level effects, the distribution of morphosyntactic categories further reflects the impact of document representation. As illustrated in Figure 4, scenarios with linearized tables exhibit elevated proportions of punctuation and symbol tokens, corresponding to layout markers and separators rather than linguistic structure.

When tables are removed, morphosyntactic profiles stabilize across parsers. The relative proportions of core categories such as nouns, verbs, and adjectives become consistent, indicating that narrative epidemiological text presents a regular grammatical structure once freed from tabular noise. This pattern holds across both general-purpose parsers and PortParser v2, reinforcing the conclusion that preprocessing choices dominate morphosyntactic behavior.

Summary. Taken together, the results demonstrate that no parser evaluated in this study is robust to the naive linearization of tables, that sentence segmentation varies substantially across parsers even for identical input text, and that differences in lemmatization, particularly in UDPipe, may affect tasks sensitive to lexical diversity. Crucially, the

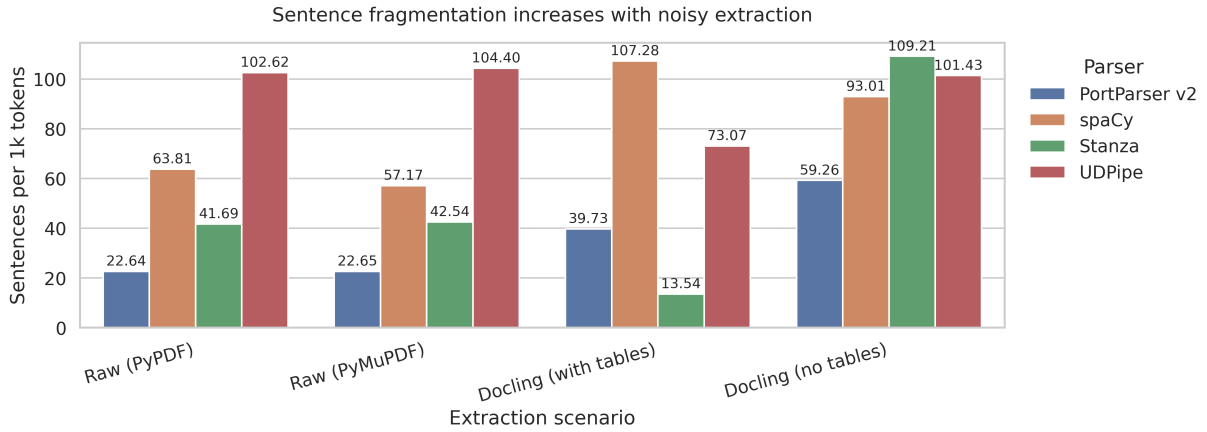


Figure 3: Sentence fragmentation across extraction scenarios, measured as sentences per 1k tokens. **Lower values indicate more stable sentence boundary detection.** Noisy extraction pipelines with linearized tables substantially increase fragmentation across all parsers, while explicit table removal (B2) yields values comparable to raw PDF extraction, suggesting more linguistically plausible segmentation.

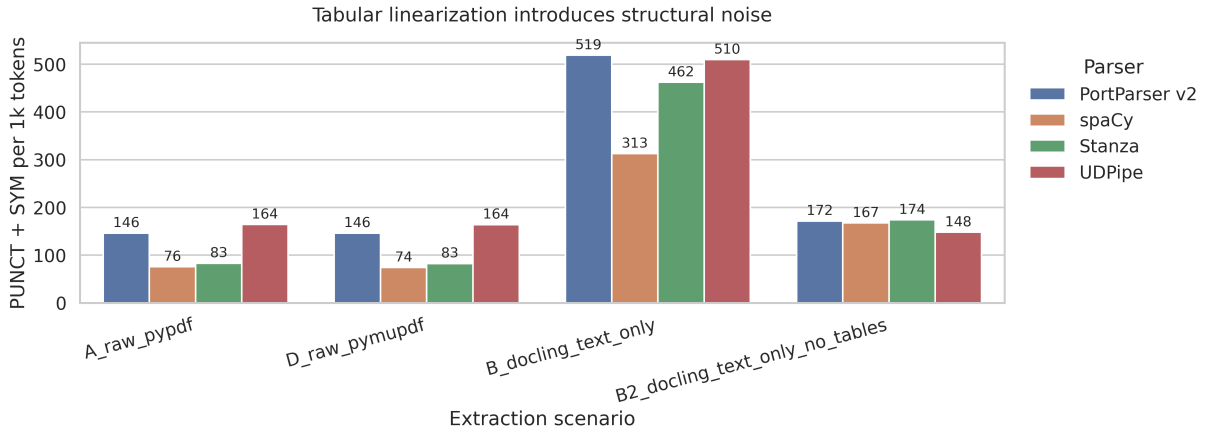


Figure 4: Structural noise across extraction scenarios and parsers, measured as the number of punctuation and symbol tokens per 1k tokens. **Lower values reflect a cleaner morphosyntactic profile with fewer non-linguistic artifacts.** Linearized tables substantially increase structural noise across all parsers, while explicit table removal (B2) yields profiles consistent with raw PDF extraction.

quantitative evidence shows that document preprocessing and representation choices are at least as critical as parser selection, underscoring the need for pipelines that explicitly preserve the distinction between narrative text and tabular data in epidemiological reports.

7 Conclusion and Future Work

This paper presented an ongoing effort to construct a Universal Dependencies corpus for Portuguese epidemiological reports derived from documents published within the Brazilian public health system. As an initial and deliberately controlled step, the analysis focused on a single, representative SIREVA-SUS report from 2024, which was used as a pilot document to evaluate the complete document extraction and parsing pipeline. Code and resources from this paper are available at the project repository[2].

The results show that document extraction and representation choices exert a strong influence on syntactic parsing behavior, even when analysis is restricted to a single report. In particular, raw PDF extraction and the linearization of tabular content lead to substantial inflation in sentence counts and structural artifacts across all parsers considered. In contrast, layout-aware extraction combined with explicit table removal produces more stable and linguistically plausible inputs, reducing segmentation noise and yielding more consistent parsing

[2] https://github.com/ChristianSF/SIREVA-SUS-Corpus

statistics. These effects are observed across both general-purpose parsers and a parser specifically designed for Portuguese, indicating that preprocessing decisions outweigh parser-specific differences in this domain, where semi-structured content is central and the original data is available exclusively in PDF format.

The comparative evaluation further highlights systematic differences in sentence segmentation and lemmatization strategies across parsers, with direct implications for downstream tasks such as information extraction, text–table alignment, and discourse-level analysis. Taken together, the findings reinforce the need for document processing pipelines that explicitly preserve the distinction between narrative text and structured data when building syntactically annotated resources from real-world technical documents.

As future work, we will extend the validated pipeline to the full collection of SIREVA-SUS epidemiological reports, moving beyond the single-document pilot to support broader generalization of the findings reported here. The complete corpus will be released together with detailed documentation and reproducible preprocessing scripts, enabling reuse for research on syntactic parsing, domain adaptation, and downstream NLP tasks in epidemiology and public health.

A key next step is the introduction of systematic quantitative evaluation. We plan to manually annotate a small but representative subset of the corpus to serve as a gold standard, enabling the computation of standard dependency parsing metrics such as Unlabeled Attachment Score (UAS) and Labeled Attachment Score (LAS). This will allow direct comparison of parser behavior in the epidemiological domain against benchmark results on general-domain Portuguese treebanks, complementing the structural indicators used in the current pilot.

Finally, future work will detail the manual annotation and curation plan for the full UD corpus. This includes defining domain-specific annotation guidelines to handle constructions that are frequent in epidemiological reports but underrepresented in existing Portuguese UD treebanks, such as abbreviated nominal phrases, table captions used as standalone sentences, and numerical expressions with embedded units and uncertainty markers.

8 Declaration of Generative AI in the writing process

During the preparation of this work, the authors used ChatGPT and Gemini in order to improve the language, grammar, and flow of the manuscript. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

References

António Branco, João Ricardo Silva, Luís Gomes, and João António Rodrigues. 2022. Universal grammatical dependencies for Portuguese with CINTIL data, LX processing and CLARIN support. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5617–5626, Marseille, France. European Language Resources Association.

Ariani Di Felippo, Norton Trevisan Roman, Thiago Alexandre Salgueiro Pardo, and Lucas Panta de Moura. 2024. The dantestocks corpus: an analysis of the distribution of universal dependencies-based part-of-speech tags. *Revista da ABRALIN*, 22(2):249–271.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spacy: Industrial-strength natural language processing in python. *Zenodo*.

IBM Research. 2024. Docling: Layout-aware document processing for complex pdfs. `https://github.com/DS4SD/docling`. Accessed: 2025-01.

Lucelene Lopes and Thiago Pardo. 2024. Towards portparser-a highly accurate parsing system for brazilian portuguese following the universal dependencies framework. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese-Vol. 1*, pages 401–410.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajic, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Portorož, Slovenia. European Language Resources Association (ELRA).

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajic, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2020. Universal dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the Twelfth International Conference on Language Resources and Evaluation (LREC 2020)*, Marseille, France. European Language Resources Association (ELRA).

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, Online. Association for Computational Linguistics.

Alexandre Rademaker, Fabricio Chalub, Livy Real, Cláudia Freitas, Eckhard Bick, and Valeria De Paiva. 2017. Universal Dependencies for Portuguese. In *Proceedings of the fourth international conference on dependency linguistics (Depling 2017)*, pages 197–206.

Magali Sanches Duran, Elvis A. de Souza, Maria das Graças Volpe Nunes, Adriana Silvina Pagano, and Thiago A. S. Pardo. 2025. Extending the enhanced Universal Dependencies – addressing subjects in pro-drop languages. In *Proceedings of the Eighth Workshop on Universal Dependencies (UDW, SyntaxFest 2025)*, pages 143–152, Ljubljana, Slovenia. Association for Computational Linguistics.

Elvis Souza and Claudia Freitas. 2023. Annotation of fixed multiword expressions (MWEs) in a Portuguese Universal Dependencies (UD) treebank: Gathering candidates from three different sources. In *Proceedings of the 2nd Edition of the Universal Dependencies Brazilian Festival*, pages 442–450, Belo Horizonte, Brazil. Association for Computational Linguistics.

Milan Straka, Jan Hajic, and Jana Straková. 2016. UDPipe: Trainable pipeline for processing CoNLL-U files. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Portorož, Slovenia. European Language Resources Association (ELRA).

Gendered Stylistic Variation in Brazilian Portuguese Google Play Reviews: A Large-Scale Study

Tiago de Melo
Amazon State University (UEA)
Manaus, Brazil
tmelo@uea.edu.br

Abstract

We study gender-associated stylistic variation in Brazilian Portuguese Google Play reviews. Using IBGE name frequencies, we infer binary gender from first names in 76.7M reviews (96 apps, 20112025), obtaining 22.25M high-confidence labels. Women-associated reviews show markedly higher paralinguistic expressivity (about 60% higher emoji density and more lengthening/punctuation), while lexical diversity (MTLD) is nearly identical across groups. Ratings are mostly positive, with men contributing relatively more 1-star reviews and women more 5-star reviews. These findings contribute to a deeper understanding of digital sociolinguistic behavior within the Brazilian context. We discuss limitations of name-based gender inference and future demographic extensions.

1 Introduction

The Google Play Store serves as a central platform where millions of users express their opinions and experiences through reviews of apps. These comments are a valuable source of feedback for developers, providing signals about bugs, feature requests, and the perceived user experience (Pagano and Maalej, 2013). However, the way feedback is communicated is not uniform and may vary according to sociodemographic factors.

Previous work in the international literature (Noei et al., 2018) suggests that gender is associated with participation and writing behavior in app stores, revealing differences in sentiment, posting frequency and topics of interest. However, most studies focus on English-language corpora and there remains a scarcity of large-scale analyzes targeted at Brazilian Portuguese. Given the lexical richness and particularities of computer-mediated communication in Brazilian digital culture (Vieira et al., 2022), investigating how men and women use written language to evaluate apps is important for a more refined understanding of local user behavior (Guedes et al., 2016; Noei and Lyons, 2022).

This paper presents a large-scale study of gender differences in Brazilian Portuguese Google Play reviews. We collected and organized a dataset with 76,695,564 reviews from 96 popular apps spanning 2011–2025. Because Google Play does not provide explicit gender metadata, we implemented a statistical inference procedure based on users' first names, using frequency data from the 2010 Brazilian Census (IBGE) as reference. This strategy yields a representative sample of 22,254,455 reviews with high-confidence gender labels (*male* or *female*) for subsequent linguistic analyzes.

Our investigation focuses on three complementary dimensions: (i) stylometric analysis, emphasizing expressivity markers and text length; (ii) lexical diversity, measured through the *Measure of Textual Lexical Diversity* (MTLD); and (iii) star-rating distributions. Consequently, we address the following research questions:

RQ1: How do writing styles differ between reviews associated with men and women, particularly regarding paralinguistic expressivity markers in Brazilian Portuguese?

RQ2: To what extent do genders differ in lexical diversity as measured by MTLD?

RQ3: Are there systematic differences in the distribution of star ratings (1–5) between reviews associated with men and women, especially at the extremes?

The results indicate that while the typical length of the review is similar between genders, female-labeled reviews are markedly more expressive, with a 60% higher emoji density and a higher incidence of vowel elongation and repeated punctuation. We also find that reviews with female marks are slightly more positive overall, concentrating on a higher

Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)
Vol. 2, pages 238–246

proportion of 5-star ratings, whereas reviews with male marks contain a higher proportion of 1-star ratings.

We make four main contributions. First, we assemble and describe a large-scale Brazilian Portuguese Google Play review dataset with transparent provenance and language filtering. Second, we provide a stylometric comparison across more than 22 million gender-labeled reviews, quantifying differences in paralinguistic expression cues. Third, we apply MTLD at scale and show that lexical diversity is virtually equivalent across genders despite robust differences in expressivity. Fourth, we characterize gendered differences in star-rating distributions, emphasizing consistent gaps at the extremes (1 and 5 stars).

2 Related Work

2.1 Name-based gender inference

Inferring gender from first names is a common strategy in large-scale observational studies, especially when gender is not available as structured metadata. Reference services and datasets typically associate names with frequency distributions by sex, enabling probabilistic assignments or threshold-based labeling. Santamaría and Mihaljevi (Santamaría and Mihaljević, 2018) compare multiple name–gender inference services and discuss performance, coverage, and biases arising from cultural and linguistic variation as well as thresholding choices. Complementarily, Mihaljevic et al. (Mihaljević et al., 2019) highlight conceptual and external-validity limitations of such inference and recommend methodological transparency, uncertainty reporting, and caution to avoid undue causal interpretations.

In the Brazilian context, IBGE's *Nomes no Brasil* project provides a name database derived from the 2010 Demographic Census, including occurrence frequencies by gender and geographic/temporal breakdowns. In its official description, IBGE reports 130,348 distinct names observed, counted separately for male and female, and clarifies that only the first name was considered (IBGE, 2016). The resource also documents practical details relevant to inference: orthographic variants are treated as distinct entries (e.g., "Ana" vs. "Anna"), diacritics are not represented, and the sex associated with records reflects what was declared at census time (IBGE, 2016). Recent work in Portuguese NLP has also explored stylometric traits in tasks such as plagiarism detection and authorship attribution, reinforcing the value of style metrics (Uka and Berger, 2024).

Unlike studies centered on academic authorship or scientific participation, this work applies name-based inference at scale to Brazilian Portuguese Google Play reviews, propagating the label inferred from the user's name to each review. Our design favors conservative labeling, explicitly preserving *ambiguous* and *unknown* cases; comparative analyzes focus on the most reliable subsets, aiming to characterize stylistic differences without implying causal interpretations.

2.2 Gender in app reviews

Software engineering and repository-mining research have used app-store reviews as signals of opinion, usability, and perceived quality, sometimes incorporating demographic slices when possible (Dąbrowski et al., 2022). In particular, Noei and Lyons (Noei and Lyons, 2022) analyze gender in Google Play reviews using name-based inference and report differences in participation and rating patterns. Such studies motivate complementary investigations in other languages and settings, as name coverage, nickname practices, and writing conventions can vary substantially across communities.

In contrast to analyzes focused on maintenance and engineering aspects, we concentrate on linguistic–stylometric and expressivity metrics in Brazilian Portuguese. We preserve uncertainty (*ambiguous*/*unknown*) and restrict comparisons to high-confidence gender labels to reduce classification bias and support reproducible interpretations.

3 Methodology

3.1 Data collection

Our dataset was collected from reviews posted on Google Play on a diverse set of apps. We selected 96 applications (apps) by popularity and restricted the corpus to Brazilian Portuguese. Language filtering relied on the Google Play metadata field `lang`, keeping only records tagged as `pt-BR`.

3.2 Gender inference

To enable gender-aware analyzes, we infer the probable gender of each review author from the `userName` field in the metadata. The procedure has three steps: (i) extraction and normalization of the name, (ii) extraction of the first-name, and

(iii) labeling into four categories: *male*, *female*, *ambiguous*, and *unknown*.

First, all `userName` values are normalized (e.g., trimming redundant spaces and standardizing case) to reduce superficial variation. We then extract the first alphabetic token of `userName` as a proxy for the first name. This first name is queried against a public IBGE-derived database[1] (2010 Demographic Census), which provides occurrence frequencies by sex, denoted `freq_f` (female) and `freq_m` (male). We rely on the 2010 Census dataset as it remains the most recent comprehensive public resource for name frequency distributions in Brazil. Although a new census was conducted in 2022, equivalent microdata of name-frequency had not been released at the time of this study.

Based on these frequencies, we define the probability of femaleness as follows:

$$p_{\text{fem}} = \frac{\texttt{freq\_f}}{\texttt{freq\_f} + \texttt{freq\_m}}.$$

Similarly, the probability of men is $p_{\text{male}} = 1 - p_{\text{fem}}$. When the first name is missing from the IBGE database, or when $\texttt{freq\_f} + \texttt{freq\_m} = 0$, p_{fem} cannot be reliably estimated.

Labeling is conservative and threshold-based. If $p_{\text{fem}} \geq 0.95$, the name is labeled as *female*; if $p_{\text{fem}} \leq 0.05$, it is labeled as *male*. For $0.05 < p_{\text{fem}} < 0.95$, the name is considered *ambiguous*, as it occurs substantially in both sexes. We assign *unknown* when gender cannot be inferred reliably, particularly for names absent from the IBGE database (nicknames, idiosyncratic spellings, non-name strings, or foreign names) or when a valid `userName` is not available.

After inference on `userName`, the label is propagated to each review, enabling gender-stratified subsets for linguistic analyzes. Name-based inference is a statistical approximation and does not represent self-declared gender identity; therefore, we keep the *ambiguous* and *unknown* categories for transparency and sensitivity analyzes.

3.3 Linguistic analyses

To characterize stylistic differences between genders, we performed three complementary analyzes. In the stylometric analysis, we compute text-format and expressivity metrics, including review length (number of tokens per review: mean, median, and 90th percentile), and paralinguistic expressivity (emoji density per 100 tokens, proportion of reviews containing at least one emoji, vowel elongation defined as repetition of the same letter three or more times, and repeated punctuation such as "!!!" or "???"), and emphasis markers such as uppercase ratio (excluding sentence-initial capitalization). We also compute the incidence of politeness markers, laughter markers, and a small inventory of informal tokens.

Our stylometric metrics are based on established work in style analysis and author profile (Koppel et al., 2002; Rangel et al., 2017), as well as studies focused on Brazilian Portuguese (Dias and Paraboni, 2020). Table 1 defines and illustrates the metrics. Politeness markers, laughter markers, and informal tokens were defined using an *ad hoc* lexicon that will be released with the dataset.

Second, lexical diversity is measured with the *Measure of Textual Lexical Diversity* (MTLD), using the threshold $\tau = 0.72$ for each gender, following the standard value in the literature (McCarthy and Jarvis, 2010). We compute MTLD by concatenating all reviews within each group (male and female), which reduces corpus-size effects in comparisons and yields a deterministic and reproducible procedure.

Third, we analyze emojis beyond overall density by examining group-specific usage patterns. We computed the relative frequency of each emoji (percentage over the total emojis within a group), repertory overlap between groups using Jaccard similarity over the sets of the 50 most frequent emojis, $J(A, B) = \frac{|A \cap B|}{|A \cup B|}$, and a functional categorization based on the taxonomy of Barbieri et al. (Barbieri et al., 2016) (e.g., *Affect/Emotion*, *Approval/Evaluation*, *Reaction/Emphasis*).

For emojis with marked differences between groups, the strength of the association is measured using the *odds ratio* (OR), defined as the ratio between the odds of emoji occurrence in the female group and in the male group. Let a be the number of occurrences of the emoji in the female group, b the number of occurrences of other emojis in the female group, c the number of occurrences of the emoji in the male group, and d the number of occurrences of other emojis in the male group. Then:

$$\text{OR} = \frac{a/b}{c/d} = \frac{ad}{bc}.$$

We compute 95% confidence intervals for OR using Woolf's approximation (Woolf, 1955) in log

[1] https://brasil.io/dataset/genero-nomes

Table 1: Definition and computation of stylometric metrics.

Metric	Computation example	Description
Mean tokens	Reviews: "*Gostei muito!*" (3 tokens) and "*Não curti o filme.*" (4 tokens). Mean = (3 + 4) / 2 = 3.5.	Total number of tokens divided by the number of reviews.
Median tokens	Token counts: 2, 3, 3, 10, 15. Median = 3.	Middle value after sorting by token count.
90th percentile tokens (P90)	If 90% of reviews have up to 20 tokens, then P90 = 20.	Token-count threshold below which 90% of reviews fall.
Emojis per 100 tokens	Review: "*Amei este app* 😍 😍" (2 emojis, 3 tokens). Thus, $2/3 \times 100 = 67$.	Total emojis divided by total tokens, multiplied by 100.
Reviews with emoji (%)	3 reviews with emojis out of 10 reviews = 30%.	Percentage of reviews that contain at least one emoji.
! per 100 tokens	–	Number of "!" per 100 tokens.
? per 100 tokens	–	Number of "?" per 100 tokens.
Uppercase ratio	5 uppercase letters / 10 letters = 0.5.	Uppercase letters divided by total letters.
Repeated punctuation (%)	Reviews containing "!!!" or "???".	Percentage of reviews with repeated punctuation.
Elongation (%)	Tokens such as "oooi" and "genteeee".	Percentage of reviews containing elongated letters.
Politeness markers (%)	"*por favor*" and "*obrigada*".	Percentage of reviews containing courtesy expressions.
Laughter markers (%)	"kkk", "haha", "rsrs".	Percentage of reviews containing laughter expressions.
Informal tokens (mean)	"tbm", "vc", "pq", "blz".	Average number of informal tokens per review.

space:

$$\mathrm{CI}_{95\%} = \exp\left(\ln(\mathrm{OR}) \pm 1.96 \times \sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}\right).$$

Finally, we compare star-rating distributions (1–5) between genders using within-group proportions, mean rating per group, and percentage-point differences, focusing on extremes (1 and 5 stars). All analyzes in this section are restricted to high-confidence gender labels (male and female), excluding *ambiguous*, *unknown*, and records without `userName`, as described in Section 3.2.

4 Results

4.1 Dataset

Our dataset comprises 76,695,564 reviews spanning 2011–2025. For gender-aware analyzes, we apply the inference procedure described in Section 3.2, which assigns a gender label to each `userName` and propagates it to the corresponding reviews. This procedure enables us to characterize (i) the population of user names (number of distinct `userName` values and first-name diversity) and (ii) the distribution of reviews by gender. We follow best practices in corpus-construction regarding provenance tracking and typological organization (Sturzeneker et al., 2022).

Table 2 summarizes the counts. In this work, "UserNames" refers to the number of distinct user names labeled by the procedure, "Unique first names" is the number of distinct first names in each category, and "Reviews" is the total number of reviews associated with each category after label propagation from `userName`. There are also reviews with missing `userName`, for which inference does not apply; we report those separately.

All analyzes in this paper consider only the 22,254,455 reviews for which we can assign high-confidence male or female labels (12,917,778 male and 9,336,677 female), excluding the ambiguous and unknown groups and records without `userName`. Although this subset represents only part of the overall crawl, more than 22 million confidently labeled reviews provide a highly representative sample for large-scale linguistic analyzes.

Table 2: Dataset summary and distribution by inferred gender.

Category	UserNames	UserNames (%)	Unique first names	Reviews	Reviews (%)
male	5,015,453	40.30	30,545	12,917,778	16.84
female	3,769,798	30.29	35,718	9,336,677	12.17
ambiguous	1,766,475	14.19	6,001	3,141,146	4.10
unknown	1,894,009	15.22	712,888	36,325,898	47.36
no `userName`	–	–	–	14,974,064	19.52
Total	12,445,735	100.00	785,152	76,695,564	100.00

4.2 Stylometric analysis

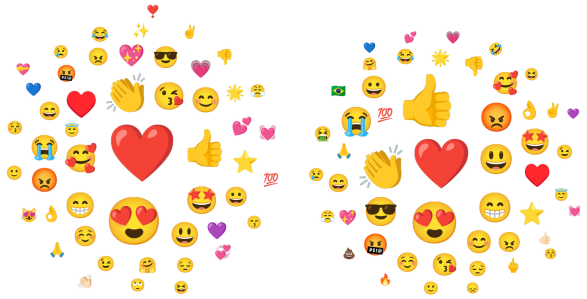
To address RQ1, we report a stylometric analysis of reviews associated with the *male* and *female* groups. Table 3 summarizes the overall differences across metrics. Regarding length, female reviews are slightly longer on average (9.54 vs. 8.89 tokens), while the median is identical (4 tokens) and the 90th percentile is very close (24 vs. 23). This pattern suggests that the length gap is not driven by the typical review (short in both groups), but by a modest and consistent increase in the upper tail, i.e., among longer reviews.

The clearest contrast emerges in the expressivity markers. Female-labeled reviews have a higher emoji density (11.57 vs. 7.17 per 100 tokens) and a higher proportion of reviews with at least one emoji (14.52 vs. 8.96%), with relative gaps above 60%. This aligns with a more expressive style, also visible in the higher incidence of elongation (5.28% vs. 3.04%) and repeated punctuation (2.97% vs. 2.34%), which are commonly associated with emphasis and pragmatic intensification in computer-mediated communication. Male-labeled reviews show a slightly higher uppercase ratio, indicating a marginal preference for all-caps emphasis. The rate of question marks per 100 tokens is very similar across groups, suggesting that question usage is not a discriminative factor in this dataset.

Although we observe small differences in politeness and laughter markers, these effects are low in magnitude and should be interpreted cautiously. Overall, our results indicate that the most robust gender-associated differences concentrate on paralinguistic and intensification cues (emojis, elongation, repeated punctuation), while length-related measures vary only modestly. These findings support the hypothesis that, in Google Play reviews, stylistic variation by gender manifests more strongly in expressivity than in content quantity, motivating finer-grained linguistic analyzes (e.g., lexical and syntactic patterns) in future work.

4.3 Emoji usage by gender

Complementing the stylometric analysis in Section 4.2 (RQ1), we examine the patterns of usage of emoji. Figure 1 shows word clouds for the 50 most frequent emojis in each group.



(a) Female (b) Male

Figure 1: Most frequent emoji clouds for female- and male-labeled reviews. Emoji size is proportional to its frequency within the corresponding group.

We observe substantial overlap in both groups' emoji repertoires, dominated by positive-valence symbols. Nevertheless, quantitative analyzes reveal systematic differences in relative frequencies and composition among the most used emojis. Table 4 lists the top five emojis in each group, with their percentages over the group's total emoji count.

The overlap between the top-50 emojis (by absolute frequency) was measured using Jaccard similarity, yielding $J = 0.82$ (approximately 45 emojis in common). This indicates substantial repertory overlap. However, usage proportions differ significantly. For example, ❤️ is 1.71 times more frequent in the female group (OR = 1.71; 95% CI: [1.70–1.72]), while 👍 is 1.84 times more frequent in the male group (OR = 1.84; 95% CI: [1.83–1.85]).

For a finer analysis, we classify emojis using a functional taxonomy adapted from Barbieri et al. (Barbieri et al., 2016). Among the 20 most frequent emojis in each group, 68% in the female group belong to the Affect/Emotion category (e.g.,

Table 3: Stylometric metrics for reviews labeled as male and female.

Metric	Male	Female	Δ (F–M)	Δ%
N (reviews)	12.917.717	9.336.646	–	–
Mean tokens	8.89	9.54	0.65	7.35
Median tokens	4	4	0	0.00
P90 tokens	23	24	1	4.35
Emojis per 100 tokens	7.17	11.57	4.40	61.35
Reviews with emoji (%)	8.96	14.52	5.55	61.95
! per 100 tokens	3.52	3.97	0.45	12.74
? per 100 tokens	0.204	0.195	-0.009	-4.37
Uppercase ratio	0.0220	0.0190	-0.0030	-13.72
Repeated punctuation (%)	2.34	2.97	0.63	26.90
Elongation (%)	3.04	5.28	2.23	73.28
Politeness markers (%)	2.07	2.20	0.13	6.45
Laughter markers (%)	0.63	0.67	0.04	6.53
Informal tokens (mean)	0.0977	0.1105	0.0128	13.10

Table 4: Top 5 emojis by gender (percentage over total emojis within the group).

Male			**Female**		
Pos.	Emoji	% (M)	Pos.	Emoji	% (F)
1	❤️	7.89	1	❤️	12.76
2	👍	7.85	2	😍	8.60
3	😍	5.53	3	👋	4.51
4	👋	4.95	4	👍	4.43
5	😃	2.87	5	🥰	2.60

❤️, 😍, 🥰, 😘), compared to 40% in the male group. The male group, in turn, uses proportionally more Approval/Evaluation emojis (👍, 👋, ⭐) and Reaction/Emphasis emojis (😡, 😎, 🔥). Table 5 reports the emojis with the largest percentage-point differences between groups.

Table 5: Emojis with the largest percentage-point differences between genders.

Emoji	Category	% Male	% Female	Δ (p.p.)
❤️	Affect/Emotion	7.89	12.76	+4.87
😍	Affect/Emotion	5.53	8.60	+3.07
👍	Approval	7.85	4.43	–3.42
🥰	Affect/Emotion	1.53	2.60	+1.07
😘	Affect/Emotion	1.58	2.56	+0.98
😎	Reaction	2.65	1.50	–1.15
😡	Reaction/Negative	2.67	1.88	–0.79

These findings complement the expressivity patterns in Section 4.2. Differential emoji usage suggests distinct discourse strategies: while female-labeled reviews rely more on affective engagement and emotional intensification, male-labeled reviews tend to emphasize approval, evaluation, or dissatisfaction. This pattern aligns with the classic work on gender in computer-mediated communication (Savicki and Kelley, 2000) and reinforces the importance of paralinguistic resources (emojis, elongation, repeated punctuation) as part of stylistic variation.

Importantly, these are large-scale tendencies, rather than linguistic determinisms. The high repertory overlap (Jaccard = 0.82) indicates that both groups share a broad set of symbols, varying their frequency and preference according to the pragmatic context.

4.4 Lexical diversity (MTLD)

To address RQ2, we evaluated the lexical diversity using MTLD. MTLD estimates the average segment length (in tokens) needed for the type–token ratio (TTR) of successive segments to fall below a threshold τ; higher values indicate greater lexical diversity. We compute MTLD with $\tau = 0.72$ for each gender following standard practice (McCarthy and Jarvis, 2010), concatenating all tokens in each group (*male* and *female*, according to Section 3.2). This reduces corpus-size effects in comparisons, while keeping the procedure deterministic and reproducible.

Overall, MTLD values are very close across groups: $\mathrm{MTLD}_M = 72.98$ and $\mathrm{MTLD}_F = 72.13$ ($\Delta = -0.85$, i.e., -1.17% for females relative to males), despite the large number of tokens analyzed (115.9M for males and 89.8M for females). Practically, this suggests that the average lexical diversity is broadly similar across genders, with

only a marginal advantage for the male group under this metric. Figure 2 shows the yearly stratification (2020–2025), the period that concentrates most gender-inferred reviews, revealing similar temporal trajectories and a sharp decline in 2022 for both groups, followed by partial recovery.

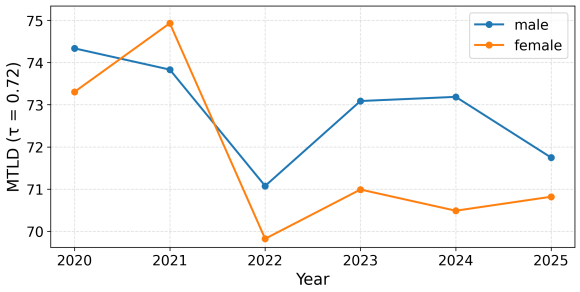


Figure 2: MTLD by year (2020–2025) for male and female groups ($\tau = 0.72$).

The 2022 dip is consistent with a shift in corpus composition (e.g., concentration of reviews in specific apps, topics, or events) and/or increased lexical conventionalization (more recurring vocabulary and formulaic evaluation patterns), which reduces diversity even without changes in typical length. As an observational slice, interpretation must be cautious. The dip may reflect factors external to gender (changes in the set of applications, user profiles, or the dynamics of the platform). Together with stylometric findings, the MTLD results indicate that clearer differences in expressivity markers (e.g., emojis and elongation) do not necessarily imply large discrepancies in lexical diversity.

4.5 Star-rating distributions

To address RQ3, we evaluated gender differences in star-rating distributions. Figure 3 compares the rating distribution (1–5 stars) between reviews labeled with male and female. For each group, we keep only reviews with a valid `score` in $\{1, 2, 3, 4, 5\}$ and a high-confidence gender label (Section 3.2). We then counted the number of reviews at each star level and normalized by the total number of reviews in that group, generating percentages within the group that sum up to 100%. In total, we analyze 12,917,778 reviews in the male group and 9,336,677 in the female group.

Both groups are strongly concentrated on positive ratings, with a predominance of 5 stars (72.07% for males and 74.21% for females). Still, the differences are consistent at the extremes: the male group has a higher fraction of 1-star ratings (13.84% vs. 10.91%, +2.93 p.p.), while the female group has a higher fraction of 5-star ratings (74.21% vs. 72.07%, +2.14 p.p.). These gaps also reflect in the mean score, which is slightly higher for females (4.316 vs. 4.213). In the intermediate range (2–4 stars), discrepancies are small, suggesting that the main distinction between groups occurs primarily between strongly negative (1 star) and strongly positive (5 stars) evaluations.

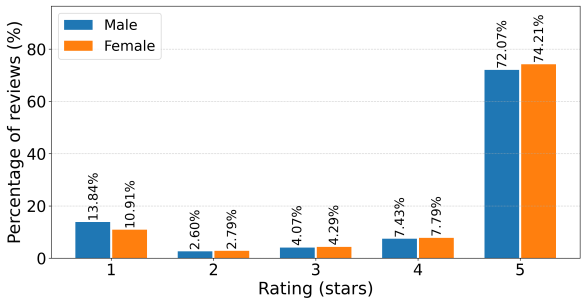


Figure 3: Star rating distribution by gender.

5 Limitations and Ethical Considerations

Interpretation should remain cautious because this is an observational study and gender is inferred from first names. Confounders such as app mix, usage profiles, and temporal effects may contribute to the observed pattern, and results should not be read as causal.

Despite the promising results, this work has limitations that warrant careful discussion, particularly regarding the ethical issues surrounding gender inference. We emphasize that gender inference in this study is a statistical approximation based on observed patterns of name-frequency. It does not capture the complexity, fluidity, or self-identification of gender (Keyes, 2018). Gender is a social and personal construct that goes beyond binary classifications or algorithmically inferred attributes. Ignoring this distinction can oversimplify identities and reinforce gender stereotypes. Therefore, any use of our analysis should be done cautiously, explicitly acknowledging these limitations.

6 Conclusions and Future Work

Our results indicate that the most consistent differences between the reviews labeled with male and female are focused on the expressivity markers. Female-labeled reviews show higher emoji density and incidence and more intensification signals, such as elongation and repeated punctuation. In contrast, global measures of content and vocabulary

are more similar. Typical review length is comparable across groups and lexical diversity measured by MTLD is virtually stable, suggesting that stylistic variation does not necessarily imply large discrepancies in lexical richness. Star ratings are mostly positive in both groups, but differ at the extremes. The male group concentrates more 1-star ratings, whereas the female group has a higher fraction of 5-star ratings and a slightly higher mean rating.

Future work follows three directions. First, improve and audit name-based gender inference by quantifying uncertainty, assessing coverage biases (e.g., names outside the reference list and nicknames), performing sample-based validations to estimate error rates, and conducting sensitivity analyzes that incorporate part of the *ambiguous* and *unknown* subsets. Second, refine the observational analysis by controlling for confounders, for example, via stratification by app/category, matching by activity period, and temporal slicing that separates composition effects (apps and usage profiles) from linguistic differences. Third, broaden the linguistic analyzes with finer-grained measurements of lexical choice and structure (e.g., semantic/affective categories, syntactic patterns, negation, and intensifiers) as well as emoji functional categories and co-occurrence with ratings, deepening the characterization of Brazilian Portuguese stylistic variation while maintaining methodological transparency about limitations.

7 Acknowledgements

The authors acknowledge the support provided by the Universidade do Estado do Amazonas (UEA) through the Academic Productivity Grant (GPA) (Administrative Ordinance No. 1177/2025-GR/UEA). This work was also supported by the National Institute of Science and Technology in Responsible Artificial Intelligence for Computational Linguistics, Information Treatment, and Dissemination (INCT-TILDIAR), funded by the Brazilian National Council for Scientific and Technological Development (CNPq), grant no. 408490/2024-1.

References

Francesco Barbieri, German Kruszewski, Francesco Ronzano, and Horacio Saggion. 2016. How cosmopolitan are emojis? exploring emojis usage and meaning over different languages with distributional semantics. In *Proceedings of the 24th ACM international conference on Multimedia*, pages 531–535.

Jacek Dąbrowski, Emmanuel Letier, Anna Perini, and Angelo Susi. 2022. Analysing app reviews for software engineering: a systematic literature review. *Empirical Software Engineering*, 27(2):43.

Rafael Dias and Ivandré Paraboni. 2020. Cross-domain author gender classification in brazilian portuguese. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1227–1234.

Gustavo Paiva Guedes, Eduardo Bezerra, Lilian Ferrari, and Fellipe Duarte. 2016. Gender differences in the use of portuguese in social networks: Evidence from liwc. In *Proceedings of the 22nd Brazilian Symposium on Multimedia and the Web*, pages 339–342.

IBGE. 2016. Um brasil de marias e josés: Ibge apresenta banco de nomes com base no censo 2010. Agência de Notícias do IBGE. Acesso em: 2026-01-13.

Os Keyes. 2018. The misgendering machines: Trans/hci implications of automatic gender recognition. *Proceedings of the ACM on human-computer interaction*, 2(CSCW):1–22.

Moshe Koppel, Shlomo Argamon, and Anat Rachel Shimoni. 2002. Automatically categorizing written texts by author gender. *Literary and linguistic computing*, 17(4):401–412.

Philip M McCarthy and Scott Jarvis. 2010. Mtld, vocd-d, and hd-d: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior research methods*, 42(2):381–392.

Helena Mihaljević, Marco Tullney, Lucía Santamaría, and Christian Steinfeldt. 2019. Reflections on gender analyses of bibliographic corpora. *Frontiers in big Data*, 2:29.

Ehsan Noei, Daniel Alencar Da Costa, and Ying Zou. 2018. Winning the app production rally. In *Proceedings of the 2018 26th ACM joint meeting on european software engineering conference and symposium on the foundations of software engineering*, pages 283–294.

Ehsan Noei and Kelly Lyons. 2022. A study of gender in user reviews on the google play store. *Empirical Software Engineering*, 27(2):34.

Dennis Pagano and Walid Maalej. 2013. User feedback in the appstore: An empirical study. In *2013 21st IEEE international requirements engineering conference (RE)*, pages 125–134. IEEE.

Francisco Rangel, Paolo Rosso, Martin Potthast, and Benno Stein. 2017. Overview of the 5th author profiling task at pan 2017: Gender and language variety identification in twitter. *Working notes papers of the CLEF*, 48.

Lucía Santamaría and Helena Mihaljević. 2018. Comparison and benchmark of name-to-gender inference services. *PeerJ Computer Science*, 4:e156.

Victor Savicki and Merle Kelley. 2000. Computer mediated communication: Gender and group composition. *CyberPsychology & Behavior*, 3(5):817–826.

Mariana Sturzeneker, Maria Clara Crespo, Maria Lina Rocha, Marcelo Finger, Maria Clara Paixão de Sousa, Vanessa Martins do Monte, and Cristiane Namiuti. 2022. Carolina's methodology: building a large corpus with provenance and typology information. In *DHandNLP@ PROPOR*, pages 53–58.

Adile Uka and Maria Berger. 2024. Could style help plagiarism detection?-a sample-based quantitative study of correlation between style specifics and plagiarism. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese-Vol. 2*, pages 103–108.

Renata Vieira, Ana Paula Banza, Ana Sofia Ribeiro, Cassia Trojahn, Fernanda Olival, Helena Cameron, Herminia Vilar, Ivo Santos, Joaquim Santos, Maria Gonçalves, and 1 others. 2022. Digital humanities and portuguese processing: a research pathway.

Barnet Woolf. 1955. On estimating the relation between blood group and disease. *Annals of human genetics*, 19(4):251–253.

A Larger Annotated Corpus of Portuguese Coreference

Evandro Fonseca
Blip
evandro.fonseca@blip.ai

Renata Vieira
University of Évora
renata.vieira@gmail.com

Abstract

Coreference resolution is a crucial task in natural language processing (NLP) that aims to identify and link expressions in a text that refer to the same entity. However, the lack of annotated data for coreference resolution in Portuguese has hindered the development of robust and accurate systems for this language. In this paper, we present an assessment of coreference annotation utilizing large language models (LLMs) for Portuguese: LLM-PREF is proposed to annotate coreference in Portuguese texts. It was evaluated and compared to a system previously proposed in the literature. The results show that although the model's world knowledge and inference capacity are quite rich - allowing it to recognize complex coreference patterns, including the pronominal anaphora phenomenon - it does not excel the previously developed rule based system.

1 Introduction

The Coreference Resolution task is challenging for Natural Language Processing, considering the required linguistic knowledge and the sophistication of language processing techniques involved. Even though it is a demanding task, a motivating factor in the study of this phenomenon is its usefulness.

Several Natural Language Processing tasks may benefit from their results, such as named entity recognition, relation extraction between named entities, summarization, sentiment analysis, among others. Coreference Resolution is a process that consists of identifying certain terms and expressions that refer to the same entity. For example, in the sentence "Joe Biden drops out of presidential race. The president says that..." we can say that [the president] is a coreference of [Joe Biden]. By grouping these referential terms, we form coreference groups, more commonly known as coreference chains.

The Coreference Resolution task for Portuguese has received some attention in past years (Coreixas, 2010; Silva, 2011; Garcia and Gamallo, 2014; Fonseca et al., 2014, 2015, 2016b; Fonseca, 2018). Most of these models were built using machine learning approaches. The most recent Portuguese coreference system, proposed by Fonseca (2018), was inspired by Lee et al. (2013).

Due to the scarcity of coreference tools for Portuguese and the recent advancements of large language models, this paper proposes the evaluation of LLMs to annotate Portuguese texts, specifically the FAPESP parallel corpus, with coreferences.

This paper is organized as follows: Section 2 presents related work on Portuguese coreference resolution; Section 3 describes the proposed annotation strategy; Section 4 presents the evaluation; Section 5 presents corpus annotation and result analysis; and finally, Section 6 presents our conclusions and future work.

2 Related Work

When dealing with the Coreference Resolution task for Portuguese, it is possible to find many corpora with some kind of coreference annotation, such as: HAREM (Freitas et al., 2010), Garcia's corpus (Garcia and Gamallo, 2014), Summ-it (Collovini et al., 2007), Summ-it++ (a new and enriched version of Summ-it (Fonseca et al., 2016a)), and Corref-PT (Fonseca et al., 2017), a semi-automatically annotated corpus.

HAREM contains annotation of named entities and their identity relations; its main purpose was the evaluation of Named Entity Recognition (NER) systems. The corpus contains manually annotated named entities distributed in ten semantic categories. Relations between these named entities have also been annotated manually, in four types: identity, inclusion, placement, and other. Garcia's corpus contains coreference annotation for person entities.

Summ-it contains noun phrase coreference an-

Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)
Vol. 2, pages 247–254

notation, being thus the corpus with the most complete coreference chains. It was semi-automatically annotated with morphosyntactic information and manually annotated with coreference. Besides coreference, the texts were also manually annotated with rhetorical relations. Also, for each text, there are manual and automatically generated summaries. Summ-it++ is an enriched version of Summ-it, adding two new annotation layers: named entities and relations between named entities. In addition, the format was changed to the SemEval (Recasens et al., 2010) format. Corref-PT (Fonseca et al., 2017) was built through a collaborative task and has 182 texts annotated semi-automatically with coreferences.

Despite the existence of many corpora, it is still difficult to produce rich coreference models using machine learning approaches due to the insufficient amount of available samples. Therefore, in this paper, we present a corpus annotated by an LLM, as an alternative for the construction of a large corpus with coreferences for Portuguese. A large corpus would allow for the production of well-trained coreference models.

Recent research on large language models (LLMs) highlights their potential and limitations in performing coreference tasks. Studies such as Gan et al. (2024) and Liu et al. (2024) emphasize the need for improvements in coreference resolution and comprehension of long contexts. Hicke and Mimno (2024) focuses on improving coreference annotation in English literary texts by exploring techniques to enhance the capabilities of LLMs in addressing the complexities inherent in literary language. Despite these efforts, challenges persist in developing rich coreference models due to limited sample availability, particularly for languages like Portuguese.

In this paper, we test whether a Portuguese coreference annotation generator based on LLMs can achieve good results without specific coreference training, comparing its performance to the latest proposed approach for Portuguese by Fonseca (2018).

3 Annotation Generation

LLM-PREF is the method proposed in this paper, based on GPT-4o (OpenAI, 2023). A detailed prompt was developed, based on a set of guidelines for annotating coreference chains. In other words, instructions to ensure that the annotation is performed while respecting the coherence and cohesion of the text, as described in Fonseca (2018). Additionally, the prompt indicates a structured JSON output, as can be seen in 3.1. The prompt was built in Portuguese, we present its original version.

3.1 Prompt

Você é um assistente que realiza anotação linguística de cadeias de correferência. Uma menção é considerada correferente de outra quando ambas se referem ao mesmo referente, ou seja, a mesma entidade, objeto ou conceito. Essa relação de correferência é importante para garantir a coesão e a coerência de um texto, pois evita repetições desnecessárias e ambiguidades. Você precisa anotar as cadeias seguindo a seguinte estrutura JSON de exemplo:

```
"chains" : [ {
   "chain_0" : [ {
     "np" : "FAPESP",
     "start" : 0,
     "end" : 5
   },
   {
     "np" : "Pesquisa FAPESP",
     "start" : 0,
     "end" : 14
   },
  ...  ]
 },
   "chain_1" : [ {
     "np" : "Lula",
     "start" : 0,
     "end" : 4
   },
   {
     "np" : "Luis inácio lula da silva"
        ,
     "start" : 100,
     "end" : 125
   },
  ...  ]
 }
 ]
}
```

As regras para que uma menção seja correferente de outra incluem:

1. ***Mesmo referente**: Como mencionado anteriormente, para que duas menções sejam correferentes, é necessário que elas se refiram ao mesmo referente. Isso significa que ambas devem se relacionar com a mesma entidade, objeto ou conceito.*

2. ***Uso de pronomes**: Uma das formas mais comuns de estabelecer a correferência entre*

duas menções é por meio do uso de pronomes. Por exemplo, se em um texto é mencionado o nome de uma pessoa do gênero masculino (referente) e, em seguida, utiliza-se o pronome "ele" para se referir a essa mesma pessoa, temos uma correferência estabelecida.

3. ***Uso de sinônimos**: Outra forma de estabelecer a correferência é por meio do uso de sinônimos. Por exemplo, se em um texto é mencionado o termo "cachorro" e, em seguida, utiliza-se o termo "animal de estimação" para se referir ao mesmo ser, temos uma correferência estabelecida.*

4. ***Proximidade textual**: A proximidade entre as menções também é um fator importante para estabelecer a correferência. Geralmente, quanto mais próximas as menções estiverem no texto, maior será a probabilidade de serem correferentes.*

5. ***Coerência e contexto**: A correferência também deve estar em conformidade com a coerência e o contexto do texto. Isso significa que as menções devem fazer sentido em relação ao restante do texto e ao tema abordado.*

6. ***Uso de conectivos**: Alguns conectivos, como "este", "aquele", "o mesmo", entre outros, também podem ser utilizados para estabelecer a correferência entre duas menções.*

7. ***Coesão**: Além de estabelecer a correferência, é importante que as menções sejam coesas, ou seja, que haja uma conexão lógica entre elas. Isso garante a fluidez e a clareza do texto.*

Em resumo, para que uma menção seja correferente de outra, é necessário que ambas se refiram ao mesmo referente, estejam próximas no texto, façam sentido em relação ao contexto e sejam coesas. O uso de pronomes, sinônimos e conectivos também pode ajudar a estabelecer essa relação de correferência. Analise o texto abaixo e anote todas suas cadeias de correferência lembrando que cada cadeia de correferência só pode ter menções de uma mesma entidade. Caso a entidade seja única, crie uma cadeia com apenas uma menção. É importante anotar todos os sintagmas nominais, nenhuma menção pode ficar para trás.

Table 1: Metrics for LLM-PREF

	MUC	**B^3**	**CEAF$_m$**	**CEAF$_e$**
Precision	62,0	50,1	57,3	32,6
Recall	29,3	21,2	32,0	25,4
F-measure	39,8	29,8	41,1	28,6
CONLL			32,7	

Table 2: BLANC metric for LLM-PREF

Coreferential Links			**Non-Coreferential Links**		
P	**R**	**F**	**P**	**R**	**F**
63.8	38.9	48.4	69.9	64.3	65.3

Table 3: Metrics for Fonseca (2018)

	MUC	**B^3**	**CEAF$_m$**	**CEAF$_e$**
Precision	53,7	50,3	49,5	44,2
Recall	52,7	47,7	53,1	57,7
F-measure	53,2	48,9	51,3	50,0
CONLL			50,7	

Table 4: BLANC metric for Fonseca (2018)

Coreferential Links			**Non-Coreferential Links**		
P	**R**	**F**	**P**	**R**	**F**
55.8	15.1	23.8	80.7	96.7	88.0

4 Evaluation

To perform the evaluation, five texts from the Summ-it corpus (Collovini et al., 2007), a manually annotated corpus with coreference, were considered. We applied five metrics from CoNLL scorer (Pradhan et al., 2014).

- MUC (Vilain et al., 1995) measures how many clusters of mentions are necessary to cover the standard chains.
- B-CUBED (Bagga and Baldwin, 1998) based on mentions, it generates results by considering the present and absent mentions of each entity in a given chain.
- CEAF (Luo, 2005) uses the alignment between entities (CEAF$_e$) or mentions (CEAF$_m$) to calculate their results
- BLANC (Recasens and Hovy, 2011) evaluates both coreference links and non-coreference links.

Each metric favors a specific feature. Additionally, we calculate CoNLL, which is the average of MUC, B^3, and CEAF_e.

The evaluation process was based on SemEval, a widely recognized corpus format for the coreference resolution task (Pradhan et al., 2011). In Table

5, the SemEval format is presented. Each line is composed of a token followed by its annotation layers (Antonitsch et al., 2016). Regarding coreferences, each noun phrase starts using "(" followed by the chain ID. Note that the ") " just occurs in the last NP token. Basically: coreferent NPs receives the same chain ID.

Table 1 presents the LLM-PREF results.

Coreference is difficult to evaluate, as we can see in the different results given by each metric. Depending on the phenomena observed, a different metric is proposed. LLM-PREF is based on the same underlying principles of Fonseca (2018), since the prompt was built upon their previous work. We can see in general higher precision and lower recall in LLM-PREF. The significantly lower recall observed in LLM-PREF compared to Fonseca (2018) is likely due to the mention detection step. While the previous rule-based system employs a dedicated and elaborate pipeline using the CoGrOO parser to explicitly extract noun phrases, our approach relies entirely on a single prompt to simultaneously identify and resolve mentions, naturally increasing the chances of missing entities. Overall, despite the higher precision, the previous rule-based system still surpasses the LLM output in terms of general metrics.

However, if we look at the BLANC metric (BiLateral Assessment of Noun Phrase Coreference), which considers coreferential and non-coreferential links separately, we see an advantage for LLM-PREF in the coreferential links. Basically, a non-coreference link is formed by two mentions that are not coreferences to each other. The BLANC metric aims to reward correct coreference chains, proportionally to their length.

Tables 3 and 4 present the results for Fonseca (2018)'s model on the same subset.

5 Corpus Annotation and Result Analysis

The Portuguese texts of the FAPESP parallel corpus (Aziz and Specia, 2011), consisting of 3,840 articles were annotated with an LLM. A total of 105 texts had to be excluded due to prompt size limitations. The resulting annotated corpus is composed of 3,735 texts, 104,348 coreference chains, and 226,973 noun phrases. Although the automatic annotation is not perfect, this resource may be useful for future research in coreference. It can be curated to produce higher quality annotations, and it can be used as a source of examples of coreference chains.

To produce the entire corpus, 18 million tokens were expended. The LLM utilized 10,416,029 tokens for input prompts, and 8,274,124 tokens for completions, resulting in a total of 18,690,153 tokens consumed. Essentially, in each iteration, the input tokens consist of the sum of prompt tokens and the text content, while the completion tokens represent the entire output, specifically the JSON output. This substantial token count underscores the extensive computational effort involved in processing the corpus.

It is important to highlight that the original FAPESP Parallel Corpus is distributed under a Creative Commons Attribution-NonCommercial (CC BY-NC 2.0) license. Therefore, the new coreference layer generated in this work inherits these non-commercial and attribution constraints, ensuring an ethical distribution and use of the corpus by the NLP research community.

Regarding the result analysis, we investigated several chains to identify the main errors in the annotation process. In chain C1, it is clear that the model mistakenly combined two separate entities, likely due to hallucination. Chain C2 presents a similar case where three distinct entities were linked.

- C1=*{[Peru], [Bolívia]}* (see context in Figure 1)

- C2=*{[a prefeitura do Rio], [o Instituto Municipal de Urbanismo Pereira Passos], [a Secretaria do Meio Ambiente]}* (see context in Figure 2)

Unfortunately, understanding the reasons behind this is quite challenging. Additionally, we identified chains that are partially correct, such as C3. In C3, it is evident that Professor Hermógenes could be considered part of the group of 105 researchers, but he does not represent the same entity.

- C3=*{[105 pesquisadores], [eles],[professor Hermógenes de Freitas Leitão]}* (see context in Figure 3)

In C4 and C5, we found a more complex case. In addition to splitting the chains of "José de Souza Martins" and "Martins," the model invented an

Table 5: SemEval annotation scheme

ID	Token	Lemma	PoS	Feat	Head	NE	Rel	Coref
1	A	o	art	F=S	_	_	_	_
2	opinião	opinião	n	F=S	0	_	_	_
3	é	ser	v-fin	PR=3S=IND	_	_	_	_
4	de	de	prp	_	_	_	_	_
5	o	o	art	M=S	_	_	_	(2
6	agrônomo	agrônomo	n	M=S	0	_	_	_
7	Miguel_Guerra	_	prop	M=S	0	PES	(9)	_
8	,		_	_	_	_	_	_
9	de	de	prp	_	_	_	_	_
10	a	o	art	F=S	_	_	_	_
11	UFSC	_	prop	F=S	0	ORG	(9)	(3)

Figure 1: Context of C1 chain

Houve um programa do governo dedicado ao crescimento econômico, com uma estratégia de incorporação, cujo lema era ``Integrar para não entregar'', lembram? Existia uma mentalidade naquele momento na América Latina de que sem uma política de ocupação dessas áreas elas seriam entregues a outros. Não só no Brasil. Os governos do **[Peru[21]]** e da **[Bolívia[21]]** tinham programas idênticos de integração nacional....

Figure 2: Context of C2 chain

Pensando nisso acabamos de fazer um trabalho pioneiro na América Latina com **[a prefeitura do Rio[10]]** com equipes de pesquisadores de geomorfologia, oceanografia, engenharia e saúde, entre outros. O objetivo foi obter um diagnóstico da vulnerabilidade potencial da cidade. O estudo foi patrocinado pel**[o Instituto Municipal de Urbanismo Pereira Passos[10]]** e chama Rio, próximos 100 anos – O aquecimento global e a cidade. Estudo semelhante foi realizado logo após para todo o estado do Rio por solicitação d**[a Secretaria do Meio Ambiente[10]]**. São trabalhos de avaliação de impactos devidos a mudanças climáticas do nível do mar.

Figure 3: Context of C3 chain

Iniciado em 1994, o ProjetoFlora Fanerogâmica do Estado de São Paulo integra **[105 pesquisadores[6]]** de 23 instituições de todo o estado na identificação das plantas com flores da região. De 1994 a 1997 **[eles[6]]** realizaram 111expedições científicas e 58 visitas ... A iniciativa do projeto Flora Fanerogâmica foi do **[professor Hermógenes de Freitas Leitão[6]]**, que morreu em fevereiro de 1996, quando dirigia uma atividade de campo da pesquisa.

additional reference of “José de Souza Martins" and missed some mentions, such as “Talcott Parsons," resulting in incorrect linkage with the pronoun “ele." Furthermore, the same chain contains an excessive number of incorrect pronominal links. It seems that the coreference task can be challenging, even for large language models.

- C4=*{[José de Souza Martins], [professor José de Souza Martins],[José de Souza Martins],[ele],[ele],...}*(see context in Figure 4)

- C5=*{[Martins], [Martins]}* (see context in Figure 4)

Despite some inconsistencies found in the last presented chains, interesting coreference chains were found, see examples C6 and C7. In these chains, we can see that the model’s world knowledge and inference capacity are quite rich, allowing it to recognize complex coreference patterns, including the pronominal anaphora phenomenon.

- C6[1]=*{[Dieter Muehe], [geógrafo baiano Dieter Carl Ernst Heino*

[1]this chain can be found in 911_3648.json file

Figure 4: Context of C4 and C5 chains

[José de Souza Martins[1]] A sociologia que examina as margens, os sonhos e a esperança Mariluce Moura e Marcos de Oliveira A escrita do **[professor José de Souza Martins[1]]** , 69 anos, é de surpreendente ... indo da própria e diferenciada trajetória pessoal desse filho de operários, **[ele[1]]** próprio um trabalhador muito precoce ... o professor **[Martins[8]]** não acredita que os sonhos sejam domínio exclusivo da psicanálise e da teoria freudiana. Em meio a tamanha riqueza de reflexões... há pouco lançado numa segunda edição pela editora Contexto. É um dos 27 livros publicados por **[Martins[8]]**, considerado por **[ele[1]]** mesmo central em sua obra sociológica ... Em tempo: **[Martins[8]]** foi professor da Cátedra Simón Bolívar na Universidade de Cambridge, na Inglaterra... Como é essa relação entre pesquisa sociológica e linguagem? - Eu aprendi sociologia no grupo... Isso era muito próprio da sociologia dos anos 1950, 60. O Talcott Parsons fez assim e era sucesso, portanto. Até o dia em que Wright Mills, outro sociólogo importante, disse que era preciso traduzir Parsons para o inglês (ora, ~~**[ele[1]]**~~ tinha escrito em inglês!). Parsons foi derrotado na revolta estudantil de 1968...

Muehe],[ele],[professor titular], [este filho de alemães nascido em Maragogipe, no Recôncavo Baian],[Ganhador do Prêmio Conrado Wessel 2003 na categoria Ciência Aplicada ao Mar, Dieter Muehe],[senhor]} (see context in Figure 5)

- C7=*{[a tuberculose],[essa doença infecto-contagiosa],[a doença]}*(see context in Figure 6)

6 Conclusion

This paper presents an alternative for annotating coreference using large language models. The motivation is based on the critical shortage of comprehensive coreference corpora for the Portuguese language, providing means for the development and enhancement of coreference resolution models. A corpus comprising 3,735 texts from the FAPESP corpus (Aziz and Specia, 2011), with 104,348 coreference chains, and 226,973 noun phrases was annotated using GPT-4o (OpenAI, 2023). GPT was used due to its availability in the work environment.

The evaluation, although limited in scope, demonstrated promising results. Despite some errors in the annotation, the overall quality of the coreference chains points to some capability of LLMs for identifying complex coreference patterns. Coreference resolution can potentially improve the performance of various natural language processing tasks such as named entity recognition, relation extraction, summarization, and sentiment analysis. This corpus can contribute to further studies on coreference.

Future work will involve expanding the evaluation to include a larger set of texts and further refining the prompts for the annotation process to minimize errors. By making Coref/FAPESP publicly available[2], we hope to foster further research and development in Portuguese coreference resolution and contribute to the broader NLP community.

References

A. Antonitsch, A. Figueira, D. Amaral, E. Fonseca, R. Vieira, and S. Collovini. 2016. Summ-it++: an enriched version of the summ-it corpus. In *Proceedings of 10th edition of the Language Resources and Evaluation Conference*, Portorož, Slovenia.

Wilker Aziz and Lucia Specia. 2011. Fully automatic compilation of a Portuguese-English parallel corpus for statistical machine translation. In *STIL 2011*, Cuiabá, MT.

Amit Bagga and Breck Baldwin. 1998. Algorithms for scoring coreference chains. In *Proceedings of the first International Conference on Language Resources and Evaluation Workshop on Linguistics Coreference*, pages 563–566, Granada, Spain.

Sandra Collovini, Thiago I Carbonel, Juliana Thiesen Fuchs, Jorge César Coelho, Lúcia Rino, and Renata Vieira. 2007. Summ-it: Um corpus anotado com informações discursivas visando a sumarização automática. In *V Workshop em Tecnologia da Informação e da Linguagem Humana*.

Tatiane Coreixas. 2010. Resolução de correferência e categorias de entidades nomeadas. Master's thesis, Pontifícia Universidade Católica Do Rio Grande Do Sul.

E. B. Fonseca, R. Vieira, and A. Vanin. 2016a. Adapting an entity centric model for portuguese coreference resolution. In *Proceedings of the 10th Annual Conference on Language Resources and Evaluation (LREC 2016)*, In Press.

[2]The corpus is available at: https://github.com/evandrofonsecatake/llm-pref

Figure 5: Context of C6 chain

Geógrafo **[Dieter Muehe[1]]** diz que é fundamental monitorar o litoral e o oceano para saber o que realmente vai mudar no clima... A resposta do **[geógrafo baiano Dieter Carl Ernst Heino Muehe[1]]** ao pedido de entrevista de Pesquisa FAPESP soou quase como um blefe. “Venha até meu apartamento, em Niterói, para conversarmos. Estou aposentado e vou pouco ao Fundão”, disse **[ele[1]]** se referindo ao o campus da Universidade Federal do Rio de Janeiro (UFRJ), onde é **[professor titular[1]]** e ainda orienta doutorandos. Aos 70 anos, a voz baixa e aparentemente tímida ao telefone dava a impressão de que se tratava de um pesquisador cansado, dedicado, a essa altura da existência, apenas a criar netos. A realidade de **[Dieter Muehe[1]]** é bem diferente do que sua discrição deixa ver. Durante a entrevista, **[este filho de alemães nascido em Maragogipe, no Recôncavo Baiano[1]]**, colocou sobre a mesa dois livros da maior importância para quem estuda a costa do país... Os dois livros tiveram a ativa participação de **[Dieter Muehe[1]]** como coordenador dos numeros... “Só com mais informações é que saberemos, nos próximos anos, o que vai realmente mudar no clima e quais as conseqüências para as populações”, alerta. **[Ganhador do Prêmio Conrado Wessel 2003 na categoria Ciência Aplicada ao Mar, Dieter Muehe[1]]** tem uma filha e dois netos.... Então teremos definitivamente fixados os nossos limites marítimos. O **[senhor[1]]** é geógrafo, mas seu trabalho abrange todas as frentes de pesquisa? Sim....

Figure 6: Context of C7 chain

Pesquisadores de Ribeirão Preto desenvolvem a primeira vacina gênica contra **[a doença[3]]** Desde que o alemão Robert Koch anunciou a descoberta do bacilo d**[a tuberculose[3]]**, em 1882, a prevenção e o tratamento d**[a doença[3]]** desafiam cientistas em todo o mundo. E pela dimensão que **[essa doença infecto-contagiosa[3]]** alcança na atualidade, torna-se particularmente importante o desenvolvimento de uma nova vacina contra **[a tuberculose[3]]** por uma equipe de pesquisadores ... considerada de terceira geração, que poderá ser aplicada no controle d**[a tuberculose[3]]**, caso seja comprovada em humanos a mesma eficiência já atestada em animais. Segundo Célio Silva, o trabalho começou em 1990, quando ele foi para Londres fazer seu ... não conferia proteção satisfatória contra **[a tuberculose[3]]**. ... o agente causador d**[a tuberculose[3]]**, se esconde dentro das células humanas e não é atingido pela ação dos anticorpos, seria necessário estimular os linfócitos T CD8, capazes de destruir especificamente as células infectadas pelos bacilos. ... agente causador d**[a doença[3]]** .... Reforçou essa posição o fato de entre um terço e metade da população mundial já estar infectada com o bacilo d**[a tuberculose[3]]**. ... Um dos problemas mais sérios relacionados com o controle d**[a tuberculose[3]]** é o aparecimento de bacilos que apresentam resistência a vários dos medicamentos utilizados no tratamento, como a isoniazida, pirazinamida, estreptomicina e rifampicina, entre outros. ... Os casos mais comuns de imunossupressão associados com **[a tuberculose[3]]** são os indivíduos com Aids, estressados, que tomam drogas imunossupressoras, alcoólatras e desnutridos, entre outros. ...ela pode até erradicar **[a tuberculose[3]]** em nosso meio", afirma o pesquisador. ... O tratamento d**[a tuberculose[3]]** feito com drogas antimicobacterianas é de longa duração - demora pelo menos seis meses.

E. B. Fonseca, R. Vieira, and A. Vanin. 2016b. Improving coreference resolution with semantic knowledge. In *Proceedings of the 12th International Conference on the Computational Processing of Portuguese (PROPOR 2016)*, In Press.

Evandro Fonseca, Vinicius Sesti, Sandra Collovini, Renata Vieira, Ana Luísa Leal, and Paulo Quaresma. 2017. Collective elaboration of a coreference annotated corpus for portuguese texts. In *Proceedings of II workshop on Evaluation of Human Language Technologies for Iberian Languages*, volume 1881, pages 68–82, Murcia, Spain.

Evandro B Fonseca, Renata Vieira, and Aline A Vanin. Resolução de correferência em língua portuguesa: Pessoa, local e organização.

Evandro B Fonseca, Renata Vieira, and Aline A Vanin. 2014. Coreference resolution in portuguese: Detecting person, location and organization. In *Journal of the Brazilian Computational Intelligence Society*, volume 12, pages 86–97.

Evandro Brasil Fonseca. 2018. Resolução de correferência nominal usando semântica em língua portuguesa.

Evandro Brasil Fonseca, Renata Vieira, and Aline Vanin. 2015. Dealing with imbalanced datasets for coreference resolution. In *The Twenty-Eighth International Flairs Conference*.

Cláudia Freitas, Cristina Mota, Diana Santos, Hugo Gonçalo Oliveira, and Paula Carvalho. 2010. Second harem: Advancing the state of the art of named entity recognition in portuguese. In *The seventh international conference on Language Resources and Evaluation*.

Yujian Gan, Juntao Yu, and Massimo Poesio. 2024. Assessing the capabilities of large language models in coreference: An evaluation. In *Joint 30th International Conference on Computational Linguistics and 14th International Conference on Language Resources and Evaluation, LREC-COLING 2024*, pages 1645–1665. European Language Resources Association (ELRA).

Marcos Garcia and Pablo Gamallo. 2014. Multilingual corpora with coreferential annotation of person entities. In *Proceedings of the 9th edition of the Language Resources and Evaluation Conference*, pages 3229–3233.

Rebecca MM Hicke and David Mimno. 2024. [lions: 1] and [tigers: 2] and [bears: 3], oh my! literary coreference annotation with llms. *arXiv preprint arXiv:2401.17922*.

Heeyoung Lee, Angel Chang, Yves Peirsman, Nathanael Chambers, Mihai Surdeanu, and Dan Jurafsky. 2013. Deterministic coreference resolution based on entity-centric, precision-ranked rules. *Computational Linguistics*, 39(4):885–916.

Yanming Liu, Xinyue Peng, Jiannan Cao, Shi Bo, Yanxin Shen, Tianyu Du, Sheng Cheng, Xun Wang, Jianwei Yin, and Xuhong Zhang. 2024. Bridging context gaps: Leveraging coreference resolution for long contextual understanding. *arXiv preprint arXiv:2410.01671*.

Xiaoqiang Luo. 2005. On coreference resolution performance metrics. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Vancouver, Canada.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Sameer Pradhan, Xiaoqiang Luo, Marta Recasens, Eduard H. Hovy, Vincent Ng, and Michael Strube. 2014. Scoring coreference partitions of predicted mentions: A reference implementation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 30–35, Baltimore, MD, USA.

Sameer Pradhan, Lance Ramshaw, Mitchell Marcus, Martha Palmer, Ralph Weischedel, and Nianwen Xue. 2011. Conll-2011 shared task: Modeling unrestricted coreference in ontonotes. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–27. Association for Computational Linguistics.

Marta Recasens and Eduard H. Hovy. 2011. BLANC: implementing the rand index for coreference evaluation. *Natural Language Engineering*, 17(4):485–510.

Marta Recasens, Lluís Màrquez, Emili Sapena, M Antònia Martí, Mariona Taulé, Véronique Hoste, Massimo Poesio, and Yannick Versley. 2010. Semeval-2010 task 1: Coreference resolution in multiple languages. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 1–8. Association for Computational Linguistics.

Jefferson Fontinele da Silva. 2011. Resolução de correferência em múltiplos documentos utilizando aprendizado não supervisionado. Master's thesis, Universidade de São Paulo.

Marc Vilain, John Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman. 1995. A model-theoretic coreference scoring scheme. In *Proceedings of the 6th Conference on Message understanding*, pages 45–52, Columbia, Maryland.

Social-RAG: A Retrieval-Augmented Generation Pipeline for Computational Social Science Research on Telegram

Leonardo Nascimento[1], **Eric Brasil**[2], **Arthur Lima**[1]
Gabriel Andrade[1], **Ricardo Sodré Andrade**[3], **Tarssio Barreto**[4]
[1] Federal University of Bahia, Salvador, Bahia, Brazil
[2] UNILAB, São Francisco do Conde, Bahia, Brazil
[3] National Archives, NE Regional Office, Salvador, Bahia, Brazil
[4] Ministry of Health, Brasília, DF, Brazil
{leofn3,profericbrasil,arthurlimareserva,gabriel.tarssioesa}@gmail.com
rsandrade@ufba.br

Abstract

Digital trace data expand empirical opportunities in the social sciences but intensify the challenge of scale: many corpora are now too large and fast-moving to read exhaustively without losing interpretive rigor. We present Social-RAG, a modular Retrieval-Augmented Generation (RAG) architecture for scalable qualitative inquiry that preserves evidence traceability, auditability, and researcher control. We apply it to public Telegram messages organized into two thematic subsets—vaccine discourse and debates on Brazil's Lei Rouanet cultural funding policy—and detail core design choices: "one post = one chunk" indexing, embedding-based semantic retrieval, an Adaptive-K cutoff for context selection, MMR re-ranking for diversity, and structured analytical instructions that constrain generation to retrieved evidence. We evaluate the system with hermeneutic and factual question blocks and compare three models (local open-weight, cloud open-weight, and commercial closed) using an LLM-as-judge protocol with qualitative criteria. Across both corpora, the larger models perform robustly on narrative and factual tasks, while the local model remains useful for exploratory narrative synthesis but is less reliable for strict factual extraction and attribution. We close with implications, limitations, and directions for improving scalability and extensibility.

1 Introduction

Digital trace data (Howison et al., 2011) and digital methods (Jungherr, 2015; Omena, 2019; Rieder and Röhle, 2017; Rogers, 2013) have introduced new empirical possibilities — and methodological tensions — into social science research (Amaturo and Aragona, 2019; Carrigan, 2014; Conte et al., 2012; Lupton, 2015; Marres, 2017; Nascimento, 2016). The massive, continuous production of data, the processes of datafication of social behavior (van Dijck, 2014; Lomborg et al., 2020; Sadowski, 2019; Southerton, 2020), and the subsequent reuse of these traces in socio-anthropological research (Salganik, 2018; Rogers, 2013) have posed renewed methodological challenges across the humanities. On the one hand, the abundance of digital traces expands the possibilities for empirical inquiry into human behavior. On the other hand, it creates a structural mismatch: the volume, velocity, and heterogeneity of the data produced often overwhelm the analytical capacity of traditional social-scientific approaches (Abbott, 2000). The scale of available data challenges established methods in the humanities, making manual pattern identification difficult. As a result, the abundance of data coexists with the risk of methodological paralysis or, worse, the uncritical adoption of tools that claim to tame the complexity of digital data traces.

These trends require re-evaluating not only data collection and processing techniques but also the epistemological framework mobilized in social research using digital data. The challenges are multiple and interconnected: they involve the critical evaluation of digital sources (Gebru et al., 2021) and their algorithmic pre-construction (boyd and Crawford, 2012; TacticalTechVideos, 2014); the articulation — rather than opposition — between qualitative and quantitative approaches; epistemological vigilance against the allure of objectivity and the power of visual evidence (Rieder and Röhle, 2017); and the fundamental distinction between data elicited by the researcher and data collected from pre-existing records (Salganik, 2018). These challenges are multifaceted, encompassing not only the need for enhanced digital literacy among researchers and sufficient computational capacity for large-scale laboratory work, but also issues related to publicity, accessibility, and representativeness.

Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)
Vol. 2, pages 255–265
April 13-16, 2026

This article [1] presents the development of Social-RAG, a Retrieval-Augmented Generation (RAG) architecture for analyzing digital trace data from Telegram groups and channels. We detail the key design decisions, technical choices, and methodological limits of this approach in humanities and social science research, situating it within broader debates on RAG as an epistemic resource for large-scale digital data analysis. We take a technical and methodological approach, describing the implemented pipeline and its practical effects on textual analysis, while avoiding black-box treatment of AI components in the research process (Schwandt, 2022).

The article has five sections: (1) a brief overview of RAG components, variants, and key debates; (2) Social-RAG and its fit to our research needs and data; (3) implementation methods (Telegram data collection, vectorization/indexing, Adaptive-K retrieval, system instructions, and the Streamlit interface); (4) evaluation (hermeneutic and factual experiments, criteria, cross-model results, limitations, and future work); and (5) a conclusion summarizing the main contributions.

2 Retrieval-Augmented Generation (RAG): definition and operating structure

The core principle of Retrieval-Augmented Generation (RAG) architectures is to retrieve relevant documents in response to a query and incorporate them into the processing context of large language models (LLMs) (Lee et al., 2025; Lewis et al., 2021). In general, a RAG system comprises two distinct components: a retrieval module, responsible for identifying and selecting pertinent passages from an external knowledge base, and a generative module, which produces answers conditioned on both the original query and the retrieved material (Zhou et al., 2023). This functional separation allows the model to rely less on knowledge internalized in its parameters during training, and instead to draw on information that is up-to-date, specialized, or situated at inference time. By grounding text generation in retrieved evidence, RAG systems tend to reduce hallucinations (fluent outputs that are nevertheless factually incorrect), increasing the traceability and verifiability of responses (Filippova, 2020; Gao et al., 2023; Maynez et al., 2020; Singh et al., 2025).

A basic RAG pipeline has three steps: index documents as embedded chunks in a vector store, retrieve semantically similar passages for an embedded query, then prompt the LLM to synthesize an answer grounded in that evidence. Output quality depends on choices such as chunking, embeddings/normalization, metadata filtering, re-ranking, and citation rules, which make the claim–source link explicit.

RAG's core value is factual grounding: it reduces hallucinations and obsolescence by forcing generation to rely on a bounded, verifiable corpus rather than opaque parametric recall. This makes outputs auditable — claims can be traced to specific passages, preserving verification practices akin to footnotes — and shifts the LLM from a "stochastic parrot" (Bender et al., 2021) toward a more controllable research instrument under the researcher's interpretive authority.

3 Implementing a Retrieval-Augmented Generation architecture: Social-RAG

RAG architectures vary widely in complexity, modularity, and methodological sophistication (Gangavarapu et al., 2025; Oche et al., 2025), and recent work has introduced iterative retrieval, knowledge structures, and tighter evidence controls to improve both retrieval and generation (Brontes et al., 2025); yet most systems still start from a common baseline that remains a useful reference point. This baseline—often called naive RAG—follows a simple Retrieve–Read paradigm (Gao et al., 2023), with single-pass retrieval (no query chaining, reassessment, or dedicated filtering) and generation without explicit strategies for interpreting or reconciling retrieved material, making it prone to weakly relevant, redundant, or contradictory passages and offering limited support for integrating multiple sources or perspectives. More broadly, RAG quality hinges on design choices across its two core axes, retrieval and generation (Zhang and Zhang, 2025): on the retrieval side, naive approaches lack systematic denoising and reasoning mechanisms to articulate lines of evidence (Cheng et al., 2025), while on the generation side, weak controls over evidence use hinder handling ambiguity, interpretive conflict, and source prioritization—central concerns in humanities and social-science research (Babbie, 2013).

[1] Esse trabalho é uma versão do artigo submetido para a revista Journal of Computational Social Sciences e publicado como preprint em https://doi.org/10.31235/osf.io/wmc2q_v1

To overcome naive RAG limits, recent approaches make the pipeline more modular and dynamic, adapting indexing, retrieval, re-ranking, and generation to task goals: Self-RAG trains the model to signal when more evidence is needed and to trigger new searches (Asai et al., 2023), Adaptive-RAG retrieves only when the current context is judged insufficient (Lee et al., 2025), other methods monitor context cutoffs (Xie et al., 2025) or learn retrieval/reordering policies via reinforcement learning (Dynamic-RAG) (Sun et al., 2025), and Agentic RAG adds explicit cycles of planning, querying, evaluation, and synthesis (Singh et al., 2025). Beyond performance gains, these designs operationalize procedures familiar to social-science inquiry — gathering relevant evidence, integrating multiple sources, and producing traceable, verifiable syntheses — so, given the demands of misinformation research and the scale, heterogeneity, and context dependence of Telegram data, we propose Social-RAG as a task-oriented architecture that balances methodological control, evidence traceability, and operational feasibility.

Social-RAG builds on core ideas from prior RAG proposals while introducing design choices tailored to humanities and social-science research needs. Its development responds to the scale and velocity of digital-platform data, which increasingly outpace traditional close-reading workflows, especially in domains such as vaccine misinformation and political extremism where rhetorical variation and platform dynamics demand methods that operate at scale without sacrificing traceability or interpretive control (Nascimento et al., 2023; Cesarino et al., 2025; Scheren et al., 2024). These corpora are also temporally volatile: messages are produced, circulated, and recontextualized in continuous flows with bursts driven by political and public health events, requiring tools that support exploratory, hypothesis-driven interpretation within short windows and sometimes near real time. We therefore conceived Social-RAG as a technical mediation layer that helps researchers navigate large corpora, surface and organize relevant evidence, and sustain auditable analytical practices.

4 Methods

Social-RAG was designed to align computational choices with corpus and common analytical workflows in the humanities and social sciences. We use a modular pipeline — vectorization, indexing, retrieval, re-ranking, and generation — so that each component can be inspected and adjusted. This approach allowed us to balance technical sophistication, methodological traceability, and operational feasibility, ensuring that the system functions as a research support instrument rather than as an opaque layer of interpretive mediation.

Social-RAG is implemented in Python using a reproducible, modular stack that integrates local and cloud components. LangChain orchestrates retrieval, context assembly, and model calls; Ollama runs open-weight models locally; and the OpenAI API is used when proprietary models are required. We use ChromaDB for persistent vector storage and embedding-based search, pandas for data ingestion and preprocessing, FastAPI as the backend service layer, and a lightweight Streamlit frontend for iterative querying and qualitative inspection, with integrated logs.

Over eight months, we followed a human-in-the-loop development cycle (Afzal et al., 2024), holding regular meetings with specialists and iteratively revising the full pipeline (dataset curation, chunking, embeddings, retrieval, and generation). Repeated team testing across datasets and query scenarios drove refinements to chunking, embedding choice, Adaptive-K retrieval, prompts and test questions, and the Streamlit interface, informed by five years of lab experience with Telegram data to assess retrieval relevance and the evidentiary quality of generated answers.

4.1 Ethics, privacy, and data governance

Following guidance in computational social science and platform research (Salganik, 2018), we treat public Telegram accessibility as insufficient to waive obligations around privacy, contextual integrity, potential harm, and researcher safety. We therefore collect only from publicly accessible groups/channels, exclude private 1:1 messages, secret chats, and closed groups, and adopt a read-only ("lurker") posture (Ferguson, 2017; Barratt and Maddox, 2016). We minimize re-identification risk by removing direct identifiers, processing user/chat references via platform IDs and/or cryptographic hashing, retaining only fields required for the analyses, and avoiding re-identification or cross-platform triangulation, mindful of "surveillance-as-method" tensions (Topinka et al., 2021). Because releasing a large, searchable corpus can amplify harm, we do not publish the full dataset (Nascimento et al.,

2023; Cesarino et al., 2025).[2] Instead, we share reproducible artifacts (evaluation materials, aggregated source data, and plotting scripts) and small anonymized samples (1,000 messages per theme) to support transparency without enabling harmful redistribution (Nascimento et al., 2023; Cesarino et al., 2025).[3]

4.2 Telegram data-collection pipeline

The data analyzed in this study comes from a computational infrastructure that automates the collection and storage of messages from public Telegram groups and channels.[4] We perform extraction via the official API (MTProto), accessed through Python libraries, and stream messages in real time for indexing in an Elasticsearch cluster together with their metadata (e.g., authorship, timestamp, content type, and forwards). After indexing, we apply transformation and enrichment routines to support querying, visualization, and analysis, with exploration via Kibana. Continuous collection is crucial because administrators or users themselves often delete content in groups and channels; nevertheless, messages already captured are preserved within the infrastructure (Ferguson, 2017).

To support analysis and experimentation, we extract thematic subsets from Elasticsearch with a script that automates querying, filtering, and export to standardized CSV/JSON, including normalization and deduplication to reduce noise from reposts and improve retrieval precision in the RAG pipeline. We evaluate the system on two contrasting subsets: vaccines (large, heterogeneous) and Lei Rouanet (smaller, more concentrated). The vaccine subset was built with wildcard queries (e.g., *vacin*, *vaccin*) over messages since 2022, yielding 116,284 unique messages after deduplication; it spans explicitly anti-vaccine posts and broader discussions of immunization, adverse effects, pharmaceutical companies, and public health policy, capturing how misinformation intersects with institutional distrust and conspiratorial narratives. The Lei Rouanet subset used orthographic variants (e.g., *rouane*, *ruane*) over messages since 2017, yielding 3,284 unique messages after deduplication, and is dominated by cultural-funding debates and "culture war" framings targeting artists and institutions.[5]

4.3 Vectorization, embeddings, and metadata

In Social-RAG, we implement a strict "**one message = one chunk**" policy: each Telegram post is indexed and retrieved as a single unit, reflecting the corpus's short, self-contained, and highly contextual discourse structure. Splitting messages risks semantic loss, while aggregating them introduces noise by mixing voices, topics, and temporalities; keeping posts intact preserves each retrieved item as an identifiable discursive act that can be cited, verified, and contextualized (Lee et al., 2025). Concretely, we embed each message with `text-embedding-3-large` (OpenAI, 2024), enabling semantic retrieval beyond keyword overlap—crucial for misinformation settings characterized by rhetorical variation, irony, abbreviations, and indirect strategies—while strengthening traceability through message-level sources with associated metadata, consistent with qualitative research practices in the humanities and social sciences (Hatch, 2010; Krippendorff, 2004).

Beyond vector representations, Social-RAG incorporates the metadata associated with each chunk. This additional layer of information is essential for analytical control, source traceability, and flexible retrieval. For each message, we store the following metadata fields: `message_id`, `chat_id`, `strict_date`, `type`, and `chat_title`.

4.4 Efficient retrieval and Adaptive-K context selection

Social-RAG uses HNSW (Hierarchical Navigable Small World) to perform low-latency nearest-neighbor search over message embeddings (Malkov and Yashunin, 2016). HNSW indexes vectors as a multi-layer graph that enables fast navigational search—coarse jumps in sparse upper layers followed by finer search in denser layers—avoiding exhaustive comparisons and supporting near-real-time exploratory querying (Malkov and Yashunin, 2016). For context construction, we avoid a fixed top-k, which is brittle across question types (Mengmeng et al., 2024): small K can miss evidence, while large K increases noise, cost, and latency

[2] Access to the full datasets is handled under controlled conditions for legitimate scholarly purposes, evaluated case-by-case and subject to commitments that prohibit re-identification and redistribution.

[3] Available after anonymous review.

[4] Messages are collected from a set of thousands of public, open groups and channels associated with far-right networks in Brazil, in real time, continuously, since 2021.

[5] Lei Rouanet (Law No. 8,313/1991) created PRONAC and federal cultural-support mechanisms, mainly via tax incentives that allow individuals and firms to allocate part of their tax liability to sponsor approved cultural projects; it also includes the National Culture Fund and other modalities. For discussion of its logic and limits, see (Balbino and Venâncio, 2020).

(Sun et al., 2025; Taguchi et al., 2025). Instead, we adopt Adaptive-K (Taguchi et al., 2025), selecting K from the similarity-score distribution in a single pass:

$$K = \underset{1 \le i < n}{\operatorname{argmax}} \left(s_{i+1} - s_i \right),$$

i.e., the sharpest drop in similarity after ranking candidates ($s_1 \ge \cdots \ge s_n$). In practice, we start from a broad candidate pool, add a small buffer ($B = 5$), restrict the cutoff search to the top 90% to avoid tail artifacts, and clamp K to $[10, 100]$ based on internal tests and corpus redundancy. Finally, we apply Maximal Marginal Relevance (MMR) to the selected set to balance relevance and diversity and reduce near-duplicate reposts (Carbonell and Goldstein, 1998).

4.5 System instructions

We developed a system prompt to guide the model's analytical behavior during Social-RAG's generation stage (the full version is available in the supplementary materials). The prompt casts the model as an analyst with interdisciplinary training in the humanities and social sciences and instructs it to respond exclusively based on retrieved passages, avoiding external knowledge and unsupported extrapolations. We adopt a "thread-of-thought" structure (Zhou et al., 2023), with explicit stages for interpreting the question, planning, selecting evidence, conducting critical analysis, and synthesis, to promote controlled, evidence-oriented outputs. We pass the corpus theme (e.g., vaccination or Lei Rouanet) as a parameter to adapt the analytical framing without changing the prompt structure. This combination strengthens Social-RAG's methodological coherence, aligns retrieval and generation with human-in-the-loop evaluation, and supports reproducibility by making the system's operational rules public.

4.6 Streamlit and the graphical interface: features and resources

Social-RAG is delivered through a Streamlit web interface (Streamlit Inc., 2021), allowing social science researchers to use the system without running scripts. Hosted on a dedicated server, the interface structures the workflow and exposes key controls, including selection of the language model, theme, and pre-vectorized corpus (e.g., Vaccine or Lei Rouanet). These choices set the analytical context and automatically load the corresponding thematic system prompt and parameters. The app supports iterative, chat-style querying with persistent history and provides access to dataset reports, the active system prompt, and a downloadable Markdown conversation log for documentation and auditing.

4.7 Models available in Social-RAG

We evaluate Social-RAG with three deliberately distinct LLM profiles to test how openness/licensing, deployment mode (local vs. cloud), and model capacity affect performance on narrative and factual tasks. Model A (`gemma3:12b`) runs locally via Ollama (Ollama, 2024a), supporting low-cost, reproducible experimentation under hardware constraints.[6] Model B (`gpt-oss:120b-cloud`) is an open-weight, large-scale cloud model accessed via Ollama (Ollama, 2024b), providing greater synthesis capacity without local infrastructure limits. Model C (`gpt-5-mini`) is a commercial closed model (OpenAI, 2025) used as a robust reference.[7]

Together, these models span common research conditions—local/control, open/scalable, and closed/optimized—enabling comparison of how cost, governance, infrastructure, and generation quality interact within the pipeline.

Unlike benchmark studies that primarily compare alternative RAG strategies (Gangavarapu et al., 2025; Zheng et al., 2023), our evaluation tests Social-RAG against the methodological and theoretical choices that guided its design. We assess performance under the corpus's analytical constraints using a human-in-the-loop procedure (Afzal et al., 2024) complemented by an LLM-as-judge protocol (Zheng et al., 2023).

Our tests use two complementary question blocks. Hermeneutic (narrative) questions evaluate whether the system can identify and synthesize recurring discursive patterns—framings, metaphors, moral judgments, and political associations—grounded in what the messages actually contain, rather than producing a single "correct" factual answer; this is central because misinformation and polarization often stabilize through narrative coherence. Factual questions evaluate precise evidence retrieval and faithful answering (enti-

[6] Ollama is a tool that allows users to download, manage, and run large language models (LLMs) through a command-line interface and a local API, facilitating both on-premise execution and, when applicable, the use of cloud backends within the same execution ecosystem. For more information, see: `https://docs.ollama.com`

[7] Through a partnership with OpenAI. OpenAI does not publicly disclose parameter counts for models in the GPT-5 family, including `gpt-5-mini`.

ties, numbers, dates, links, explicit claims), including negation and absence of information—crucial for detecting hallucinations, unwarranted extrapolations, and attribution errors in misinformation-heavy corpora.

Separating narrative from factual tasks targets distinct capabilities that automated RAG evaluations often aggregate (Zheng et al., 2023), despite their different epistemological and technical demands. Consistent with human-in-the-loop evaluation (Afzal et al., 2024), we do not rely on a closed gold standard; instead, we interpret outputs against discursive patterns established in the literature and empirical research. Applying the same protocol to two thematically distinct datasets (vaccines and Lei Rouanet) further enables comparison under variation in scale, redundancy, and the stabilization of ideological framings.

4.8 Criteria and methods for evaluating responses

We evaluated Social-RAG with task-specific qualitative criteria summarized in Table 1. For hermeneutic questions, we focused on thematic accuracy, analytical adequacy, evidence precision, synthesis, and political sensitivity; for factual questions, we prioritized verifiable extraction (question–evidence correspondence, coverage, entity extraction, sensitivity to negation/absence, and accurate recovery of numbers and links). Across both blocks, we also scored clarity/organization and the absence of hallucinations, especially corpus-unsupported gap-filling. Methodologically, this protocol approximates AI-assisted grounded theory (Charmaz, 2006; Corbin and Strauss, 2008), combining inductive pattern identification with computational retrieval and synthesis while preserving researcher interpretive control.

For comparative evaluation, we adopted an LLM-as-judge protocol to minimize order effects and label bias and to assess judge stability (Zheng et al., 2023). We conducted three blinded rounds in which the three models were rotated across positions A/B/C (each model appearing once per position), while judges saw only Response 1/2/3 with randomized order; the hidden mappings were logged, and the same two LLM judges (GPT-5 and Gemini Pro) scored all outputs under identical instructions. We replicated the same blinded procedure with a human judge (one author) using the same criteria, prompts, and response sets; Table 2 reports the blinding schedule. Across two themes (vaccine, Rouanet) and two question types (hermeneutic, factual), this yielded 12 LLM judgment tables (3 rounds × 2 themes × 2 types) plus one human table, totaling 13.

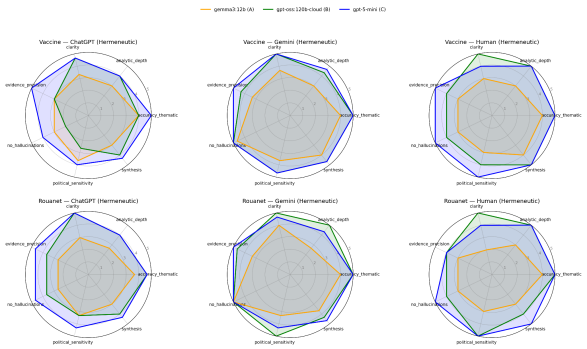


Figure 1: Hermeneutic Evaluation (by Judge): Vaccine and Rouanet — Criterion-Level Radar Profiles.

4.9 Results

All score sheets from the three blinded rounds were merged into a single long-format dataset (`judgments_long_full.csv`) with columns `theme`, `category`, `judge`, `blind`, `response_id`, `model_label`, `model_real`, `criterion`, `score`. From this file we computed criterion-level and overall mean scores by theme (vaccine/Rouanet), task type (hermeneutic/factual), model, and judge, plus judge-aggregated criterion profiles. Results are summarized with per-judge radar plots, a heatmap of overall mean variation, and a consolidated four-panel radar that aggregates criterion patterns across judges for each theme × question type.

Across both themes, hermeneutic radar profiles (Figure 1) show a stable ordering: Model C (`gpt-5-mini`) scores highest, followed by Model B (`gpt-oss:120b-cloud`), with Model A (`gemma3:12b`) trailing. This ranking is consistent across the two LLM judges and the blinded human evaluation, and is most pronounced on criteria tied to interpretive rigor (analytical depth, synthesis, clarity, and disciplined use of retrieved evidence), where Model A more often produces thinner or more generic summaries.

In the factual block (Figure 2), the radar plots show a clearer gap between the local model (A) and the larger models (B and C). Models B and C sustain consistently high scores for literal precision, coverage, entity extraction, and sensitivity to negations/absences, indicating stronger reliability when answers must be tightly anchored in explicit corpus

Question type	Criterion	Brief description
Hermeneutic	Thematic accuracy	Correspondence between the answer and the main discursive framings present in the corpus
	Analytical depth	Ability to articulate ideological, moral, and political dimensions in a non-superficial way
	Evidence precision	Appropriate and consistent use of retrieved passages to support the interpretation
	Synthesis capacity	Coherent organization of multiple discursive fragments into an intelligible explanation
	Political sensitivity	Attention to polarization contexts, avoiding undue simplifications or neutralizations
	Clarity and organization	Clear, comprehensible, and well-structured textual organization
	Absence of hallucinations	Does not introduce information, interpretations, or patterns not supported by the corpus
Factual	Literal precision	Direct correspondence between the question and retrieved textual passages
	Factual coverage	Ability to identify multiple relevant occurrences when present
	Entity extraction	Correct identification of proper names, institutions, artists, and political actors
	Sensitivity to negations	Explicit recognition of negations, absences, or contradictions in the corpus
	Number and percentage extraction	Correct retrieval of values, amounts, dates, and percentages
	URLs and links	Accurate identification and retrieval of cited links and domains
	Clarity and organization	Clear and structured presentation of factual information
	Absence of hallucinations	No fabrication of data, numbers, or sources not present in the corpus

Table 1: Evaluation criteria used in Social-RAG tests.

Blind round	Response 1	Response 2	Response 3	Judges
Blind 1	gemma3-12b	gptoss-120b	gpt5-mini	GPT-5; Gemini Pro
Blind 2	gptoss-120b	gpt5-mini	gemma3-12b	GPT-5; Gemini Pro
Blind 3	gpt5-mini	gemma3-12b	gptoss-120b	GPT-5; Gemini Pro

Table 2: Blinded evaluation schedule: model-to-response mapping across the three rounds.

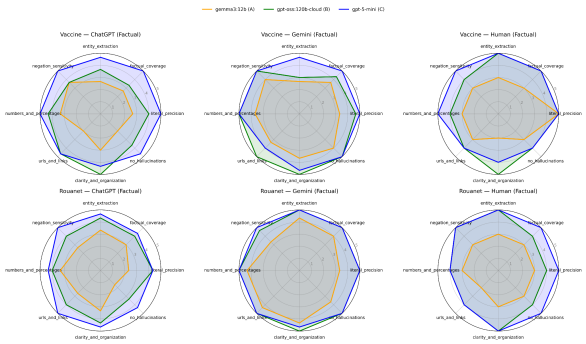


Figure 2: Factual Evaluation (by Judge): Vaccine and Rouanet — Criterion-Level Radar Profiles.

content. Model A remains usable in many cases but more often drops on precision- and evidence-dependent criteria (especially literal precision, coverage, and negation sensitivity), where small errors materially affect interpretation.

To test judge stability under the same blinding, we computed overall mean scores for each judge × subset × model (averaging across items and criteria). The heatmap (Figure 3) shows two consistent patterns: model ranking is stable across judges and subsets (C highest, B close behind, A lowest), and although judges differ in absolute severity (Gemini

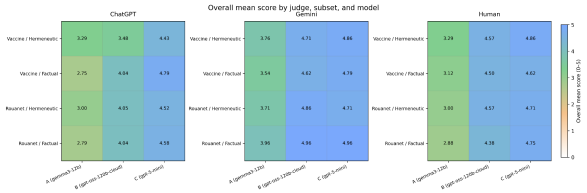


Figure 3: Overall Mean Score Variation: Judge × Subset × Model.

Pro scores higher on average than ChatGPT/GPT-5 and the human judge), the direction of differences is preserved, indicating robust comparative judgments.

Across subsets, model gaps are slightly larger in the vaccine corpus than in Rouanet, most clearly for hermeneutic questions. This likely reflects vaccines' greater volume and narrative heterogeneity, which accentuates differences in organization, inference control, and evidential discipline. In Rouanet, where framings are more stabilized and repetitive, performance remains differentiated but converges somewhat, suggesting that discursive redundancy can narrow the advantage of higher-capacity models for some interpretive tasks.

To summarize criterion-level patterns, we report consolidated radar plots by theme and question type, aggregating across judges with equal judge weighting (Figure 4). We first compute criterion means within each judge (pooling the three blinded rounds for ChatGPT/GPT-5 and Gemini Pro), then average across judges so the LLM

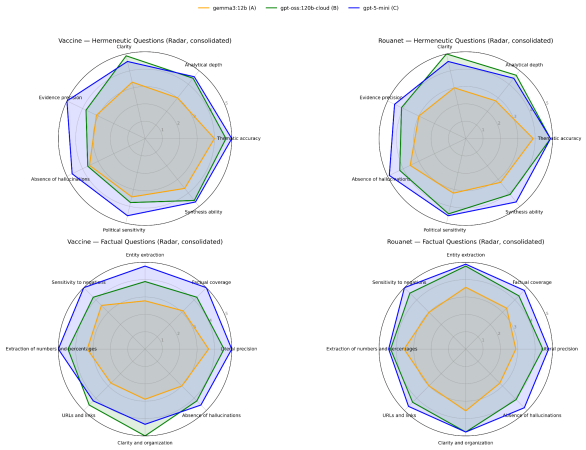


Figure 4: Consolidated Criterion Profiles (Equal Judge Weighting): Vaccine/Rouanet × Hermeneutic/Factual.

judges are not implicitly over-weighted relative to the single human evaluation. The consolidated profiles confirm the main trade-offs: in hermeneutic tasks, Models B and C maintain broad, strong profiles, while Model A is more uneven—often thematically aligned but weaker in depth, synthesis, and evidence precision; in factual tasks, Models B and C cluster at the top across most criteria, whereas Model A drops more on extraction- and verification-dependent criteria. Overall, Social-RAG yields stable comparisons across corpora and task types, and the results show that model capacity interacts with task demands: smaller local models can support exploratory interpretation, but the reliability gap widens when tasks require literal recovery, negation handling, and traceable evidence.

4.10 Limitations and future work

Social-RAG is not designed for exact counting or classical descriptive statistics (e.g., term/frequency counts, link totals), which are better handled by regex and traditional NLP pipelines. Instead, it retrieves semantically relevant messages and uses LLMs to synthesize and interpret them under explicit analytical instructions; its outputs are therefore evidence-oriented samples rather than exhaustive measurements. Social-RAG should be read as a qualitative, hermeneutic aid for exploring discursive patterns and narrative framings—supporting iterative, reflexive analysis—rather than a metric-producing system.

Future work focuses on modular evolution and scalability. Because components (embedding, indexing, retrieval, re-ranking, generation, interface) are replaceable, the system can track rapid changes in models and methods, but scaling to larger corpora and higher query concurrency will depend on compute/storage capacity and API costs. We also plan to add knowledge-graph modules to connect entities, actors, and recurring claims across messages, enabling graph-informed retrieval and more structured analysis.

5 Conclusions

This paper presented the implementation and evaluation of Social-RAG, a Retrieval-Augmented Generation architecture tailored to humanities and social-science analysis. Our starting point is that the scale and velocity of digital trace data require more than "computational power": they demand pipelines that preserve evidential traceability, interpretive control, and critical verification.

Social-RAG operationalizes this through design choices matched to Telegram-style corpora — one-post-per-chunk indexing, Adaptive-K context selection, MMR diversification, and structured analytical instructions — and we show, across two thematic datasets (vaccines and Lei Rouanet), that the system behaves consistently on both hermeneutic and factual tasks, supporting narrative synthesis and evidence recovery. Comparative experiments indicate a clear trade-off: larger models (B and C) are more reliable across both task types when evidential discipline is enforced, while the smaller local model (A) remains useful for exploratory interpretation but is less dependable for strict factual extraction, negation handling, and precise attribution. Finally, by documenting prompts, parameters, and design decisions, we make the pipeline auditable and reproducible, enabling inspection and adaptation to other corpora and research constraints.

Social-RAG neither replaces critical reading nor automates interpretation; it functions as a mediation layer that expands researchers' exploratory capacity when corpora exceeds the reach of exhaustive reading. Looking ahead, we highlight three development priorities: improving system scalability, conducting systematic evaluation across additional thematic domains, and integrating knowledge graphs to enrich contextualization. Ultimately, this work contributes to the development of computational infrastructures that serve social research rather than black-box systems that substitute for it.

Conflict of Interest

On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

Andrew Abbott. 2000. Reflections on the future of sociology. *Contemporary Sociology*, 29(2):296.

Asad Afzal, Ashwin Kowsik, Reza Fani, and Florian Matthes. 2024. Towards optimizing and evaluating a retrieval augmented QA chatbot using LLMs with human in the loop. *Preprint*, arXiv:2407.05925.

Enrica Amaturo and Biagio Aragona. 2019. Per un'epistemologia del digitale: Note sull'uso di big data e computazione nella ricerca sociale. *Quaderni di Sociologia*, 81.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2023. Self-RAG: Self-reflective retrieval augmented generation.

Earl R. Babbie. 2013. *The Practice of Social Research*, 13 edition. Wadsworth, Cengage Learning.

Gustavo Matias Soares Balbino and Renato Pinto Venâncio. 2020. Políticas culturais e arquivos públicos: o caso da Lei Rouanet. *ÁGORA: Arquivologia em debate*, 30(60):57–74.

Monica J. Barratt and Alexia Maddox. 2016. Active engagement with stigmatised communities through digital ethnography. *Qualitative Research*, 16(6):701–719.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, pages 610–623.

danah boyd and Kate Crawford. 2012. Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society*, 15(5):662–679.

Flavius Brontes, Janus Genesis, Zephyrine Noa, and Stavros Nymphodoros. 2025. Learning to retrieve, generate, and compress: A unified view of efficient RAG.

Jaime Carbonell and Jade Goldstein. 1998. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 335–336.

Mark Carrigan. 2014. An agenda for digital sociology.

Letícia Cesarino, Leonardo Nascimento, and Priscila Fonseca. 2025. Democracy "inside out": On far-right refracted publics in Brazil. In Zizi Papacharissi, editor, *The Routledge Companion to Digital Media and Democracy*, page 490. Routledge.

Kathy Charmaz. 2006. *Constructing Grounded Theory*. Sage Publications.

Ming Cheng, Yang Luo, Jianguo Ouyang, and 1 others. 2025. A survey on knowledge-oriented retrieval-augmented generation. *Preprint*, arXiv:2503.10677.

Rosaria Conte, Nigel Gilbert, Guido Bonelli, and 1 others. 2012. Manifesto of computational social science. *The European Physical Journal Special Topics*, 214(1):325–346.

Juliet Corbin and Anselm Strauss. 2008. *Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory*, 3 edition. SAGE Publications.

Ross-Helen Ferguson. 2017. Offline 'stranger' and online lurker: Methods for an ethnography of illicit transactions on the darknet. *Qualitative Research*, 17(6):683–698.

Katja Filippova. 2020. Controlled hallucinations: Learning to generate faithfully from noisy data. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 864–870.

Rohith Gangavarapu, Abhinav R. A. Srinivasan, and Venkatesh Moparthi. 2025. Evaluating accuracy in large language models: Benchmarking corrective RAG vs. naive retrieval augmented generation approach. In *2025 IEEE International Conference on AI and Data Analytics (ICAD)*, pages 1–7.

Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. 2023. Precise zero-shot dense retrieval without relevance labels. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1762–1777.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, and 1 others. 2021. Datasheets for datasets. *Preprint*, arXiv:1803.09010.

J. Amos Hatch. 2010. *Doing Qualitative Research in Education Settings*. SUNY Press.

James Howison, Andrea Wiggins, and Kevin Crowston. 2011. Validity issues in the use of social network analysis with digital trace data. *Journal of the Association for Information Systems*, 12(12).

Andreas Jungherr. 2015. *Analyzing Political Communication with Digital Trace Data: The Role of Twitter Messages in Social Science Research*. Springer.

Klaus H. Krippendorff. 2004. *Content Analysis: An Introduction to Its Methodology*, 2 edition. Sage Publications.

Jong Hyuk Lee, Ghulam Ali, and Jeong-In Hwang. 2025. A retrieval-augmented generation system for accurate and contextual historical analysis. *Computer Animation and Virtual Worlds*, 36(4):e70048.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, and 1 others. 2021. Retrieval-augmented generation for knowledge-intensive NLP tasks. *Preprint*, arXiv:2005.11401.

Stine Lomborg, Lina Dencik, and Hallvard Moe. 2020. Methods for datafication, datafication of methods: Introduction to the special issue. *European Journal of Communication*.

Deborah Lupton. 2015. *Digital Sociology*. Routledge.

Yury A. Malkov and Dmitry A. Yashunin. 2016. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *Preprint*, arXiv:1603.09320.

Noortje Marres. 2017. *Digital Sociology: The Reinvention of Social Research*. Polity Press.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919.

Song Mengmeng, Liu Zhibin, Wang Qingwei, Huang Man, and Xu Feiyang. 2024. An effective retrieval method to improve RAG performance. In *2024 7th International Conference on Data Science and Information Technology (DSIT)*, pages 1–5.

Leonardo F. Nascimento, Taciana Barreto, Leticia Cesarino, Vânia Mussa, and Priscila Fonseca. 2023. Públicos refratados: Grupos de extrema-direita brasileiros na plataforma Telegram. *Internet & Sociedade*.

Leonardo Fernandes Nascimento. 2016. A sociologia digital: Um desafio para o século XXI. *Sociologias*, 18:216–241.

Akhere Josephine Oche, Adegboyega G. Folashade, Tirthankar Ghosal, and Arindam Biswas. 2025. A systematic review of key retrieval-augmented generation (RAG) systems: Progress, gaps, and future directions. *Preprint*, arXiv:2507.18910.

Ollama. 2024a. Gemma3.

Ollama. 2024b. gpt-oss.

Janna Joceli Omena. 2019. *Métodos Digitais: Teoria-Prática-Crítica*. ICNOVA.

OpenAI. 2024. text-embedding-3-large model.

OpenAI. 2025. GPT-5 mini model.

Bernhard Rieder and Theo Röhle. 2017. Digital methods. In Mirko Tobias Schäfer and Karin van Es, editors, *The Datafied Society*, pages 109–124. Amsterdam University Press.

Richard Rogers. 2013. *Digital Methods*. MIT Press.

Jathan Sadowski. 2019. When data is capital: Datafication, accumulation, and extraction. *Big Data & Society*, 6(1).

Matthew J. Salganik. 2018. *Bit by Bit: Social Research in the Digital Age*. Princeton University Press.

Marcos L. Scheren, Vinícius S. Rodrigues, Guilherme D. López Zamora, and 1 others. 2024. Métodos mistos para a antropologia digital: Um relato de experiência sobre a análise de grupos bolsonaristas na plataforma Telegram. *Horizontes Antropológicos*, 30:e680407.

Silke Schwandt. 2022. Opening the black box of interpretation: Digital history practices as models of knowledge. *History and Theory*, 61(4):77–85.

Aditi Singh, Abul Ehtesham, Saket Kumar, and Tala Talaei Khoei. 2025. Agentic retrieval-augmented generation: A survey on agentic RAG. *Preprint*, arXiv:2501.09136.

Clare Southerton. 2020. Datafication. In Laurie A. Schintler and Connie L. McNeely, editors, *Encyclopedia of Big Data*, pages 1–4. Springer International Publishing.

Streamlit Inc. 2021. Streamlit — a faster way to build and share data apps.

Jiajie Sun, Xin Zhong, Shirui Zhou, and Jiawei Han. 2025. DynamicRAG: Leveraging outputs of large language model as feedback for dynamic reranking in retrieval-augmented generation. *Preprint*, arXiv:2505.07233.

TacticalTechVideos. 2014. Smari McCarthy, making data speak.

Chihiro Taguchi, Saku Maekawa, and Nikita Bhutani. 2025. Efficient context selection for long-context QA: No tuning, no iteration, just adaptive-k. *Preprint*, arXiv:2506.08479.

Robert Topinka, Alan Finlayson, and Camille Osborne-Carey. 2021. The trap of tracking: Digital methods, surveillance, and the far right. *Surveillance & Society*, 19(3):384–388.

José van Dijck. 2014. Datafication, dataism and dataveillance: Big Data between scientific paradigm and ideology. *Surveillance & Society*, 12(2):197–208.

Ruijie Xie, Junxiong Wang, Paul Rosu, and 1 others. 2025. Language models (mostly) know when to stop reading. *Preprint*, arXiv:2502.01025.

Wei Zhang and Jing Zhang. 2025. Hallucination mitigation for retrieval-augmented large language models: A review. *Mathematics*, 13(5).

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, and 1 others. 2023. Judging LLM-as-a-judge with MT-Bench and chatbot arena. *Preprint*, arXiv:2306.05685.

Yucheng Zhou, Xiubo Geng, Tao Shen, and 1 others. 2023. Thread of thought unraveling chaotic contexts. *Preprint*, arXiv:2311.08734.

Comida e bebida nas literaturas portuguesa e brasileira: o projeto *ReadingFood*

Diana Santos
Linguateca & Univ. of Oslo
Postboks 1003 Blindern
N-0315 Oslo, Noruega
d.s.m.santos@ilos.uio.no

Eckhard Bick
Univ. of Southern Denmark
Campusvej 55
DK-5230 Odense M, Dinamarca
eckhard.bick@gmail.com

Cristina Mota
Linguateca & INESC
R. Alves Redol 9
1000-029 Lisboa, Portugal
cristina.mota@inesc-id.pt

Resumo

Neste artigo descrevemos brevemente o projeto *ReadingFood* sobre o campo semântico da comida e bebida na literatura, que pretende comparar as obras de quatro países no período 1840-1920, mas cingindo-nos a Portugal e ao Brasil. Após apresentar as infraestruturas já desenvolvidas, tornando pública a pesquisa neste domínio, apresentamos o trabalho já feito e alguns estudos preliminares: a criação de uma taxonomia do domínio na literatura, a desambiguação em contexto, e o estudo de refeições (ou eventos relacionados com comida e bebida).

1 Motivação

Desde os anos 90 do século passado que surgiu a leitura distante, uma forma de complementar os estudos literários com métodos computacionais de maior abrangências, que não necessitassem da leitura próxima de milhares ou mesmo milhões de obras, veja-se Moretti (2013). Dessa forma surgiram trabalhos dos mais variados géneros em linguística com corpos, em que os corpos eram constituídos por obras literárias.

Para o português existem infelizmente menos obras digitalizadas e acessíveis do que para o inglês, mas pelo menos as bibliotecas nacionais de Portugal e do Brasil, e projetos como a Biblioteca Digital de Literatura de Países Lusófonos[1], o projeto Gutenberg[2], o projeto LusoLivros[3], o projeto Adamastor[4] e o Internet Archive[5], digitalizaram (e, em alguns casos, modernizaram) centenas ou mesmo milhares de obras em português. Veja-se Schöch et al. (2021) para mais informação de como coligir uma coleção de obras de Portugal, e algumas das diferenças entre estes projetos.

Depois existem projetos que tentam enriquecer essas obras em termos linguísticos, anotando-as e tornando a procura no conjunto de obras acessível através da internet. Um exemplo é a Literateca (Santos, 2019), que associa às obras tanto informação linguística (morfossintática e semântica) sobre cada palavra (Santos, 2014), como informação que podemos considerar como informação literária, por exemplo as personagens (Santos e Freitas, 2019; Santos et al., 2026a).

No projeto em apreço, ReadingFood (Santos et al.), associamos aos textos literários informação sobre comida e bebida, que reputamos um domínio fundamental para estudar a cultura e a literatura (Meigs, 1997; Boyce e Fitzpatrick, 2017; Coghlan, 2020; Barkan, 2021). Neste projeto, em curso, além de identificar a presença do domínio e os vários usos em texto literário em diferentes autores e escolas em quatro literaturas (portuguesa, brasileira, norueguesa e italiana), pretendemos:

- investigar a presença de metáforas e expressões idiomáticas;
- identificar expressões convencionais associadas a bebida e comida (rituais de convite, de brinde, de início e fim de refeição, por exemplo), como elogiar ou descrever comida (Traverso e Dimachki, 2017) e/ou pessoas (Korthals, 2008);
- localizar cenas associadas a refeições ou à sua preparação que façam parte do enredo (Schank e Abelson, 1977), estudando também a sua função (Brown, 1984);
- comparar as quatro culturas envolvidas segundo todos estas vertentes.

No presente artigo, referimo-nos simplesmente ao português, e apresentamos as primeiras duas tarefas: a classificação das palavras na secção 2 e a sua desambiguação em contexto na secção 3, seguida

[1] https://literaturabrasileira.ufsc.br
[2] https://www.gutenberg.org/
[3] aparentemente descontinuado
[4] https://projectoadamastor.org/
[5] https://archive.org/

Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)
Vol. 2, pages 266–274

de alguns estudos preliminares sobre a anotação e classificação de refeições, na secção 4.

2 Taxonomia

Começámos por desenvolver uma taxonomia dos conceitos que nos pareceram estar associados às palavras de comida e bebida que encontrámos no nosso corpo de obras literárias, formado essencialmente por obras do século XIX e princípio do século XX. Os números referem-se ao número de palavras distintas já classificadas nessa categoria. Não sendo definitiva, apresentamo-la na Figura 1.

food-gen (31) palavras gerais que se referem a comida

food-h (200) palavras que se referem a uma comida medeada por um processo qualquer culinário, e que têm um nome

food-intermediate (54) palavras que se referem a ingredientes ou especiarias, mas que não são comidas separadamente

food-edibleplant (145) palavras que descrevem entidades do mundo vegetal que são comidas (cozinhadas ou cruas)

food-plantpart (3) palavras que descrevem partes de uma entidade do mundo vegetal comestível

food-edibleanimal (83) palavras que se referem a animais que são comidos

food-ediblebodypart (46) palavras que se referem a partes de animais que se comem

food-part (33) palavras usadas para definir uma parte de algo comestível

drink-part (10) palavras usadas para referir uma parte de bebida

food-meth (17) palavras usadas para descrever um método de tratar os alimentos

food-class (39) palavras usadas para descrever uma classe de alimentos

con-food (42) palavras que descrevem um contentor de comida

con-drink (39) palavras que descrevem contentores de bebida

tool-foodprepare (2) utensílios de cozinha

tool-eat (4) utensílios para comer

Hprof-food (18) profissões associadas a comida

Lh-foodprepare (10) locais de preparação ou armazenamento de comida

Lh-eat (5) locais onde se come

Lh-drink (7) locais onde se bebe

drink (9) bebidas em geral, alcoólicas ou não

drink-non-alco (31) bebidas não alcoólicas

drink-alco (126) bebidas alcoólicas

occ-eat (23) palavras que se referem a ocasiões em que se come

occ-drink (5) palavras que se referem a ocasiões em que se bebe

jfood (55) palavras que descrevem qualidades da comida ou bebida

Hfood (4) pessoas descritas em relação à sua atitude à comida

Hdrink (7) pessoas descritas em relação à sua atitude à bebida

Como será evidente, em muitos casos uma mesma palavra pertence a mais do que uma categoria, e é por isso que procedemos à desambiguação em contexto.

Assim, muitas palavras descrevem tanto uma quantidade de bebida como um contentor para bebidas: *copo*, *xícara*, *chávena*, *cálice*, ... assim como muitas palavras descrevem (fora do contexto) uma refeição, ou aquilo que se comeu ou bebeu à refeição: *jantar*, *café*... Também alguns nomes de "tratamentos" dados à comida ou à classe de comida são usados para referir aquilo que se comeu ou mesmo nomes de receitas: *assado*, *cozido*, *sopa*...

Além disso, é evidente que a maior parte das palavras do campo da comida e bebida não pertencem exclusivamente a esse campo, quer por homonímia, como *copo* parte de espada, ou *rancho* edifício, como pelas várias relações de extensão de sentido ou metafóricas: a) de comida para outros domínios, como *coco* para chapéu de coco, ou *apimentar as conversações*, ou b) doutros domínios para comida, como *ladrilhos* de marmelada, ou c) ainda casos mais gerais que se aplicam a vários domínios, um

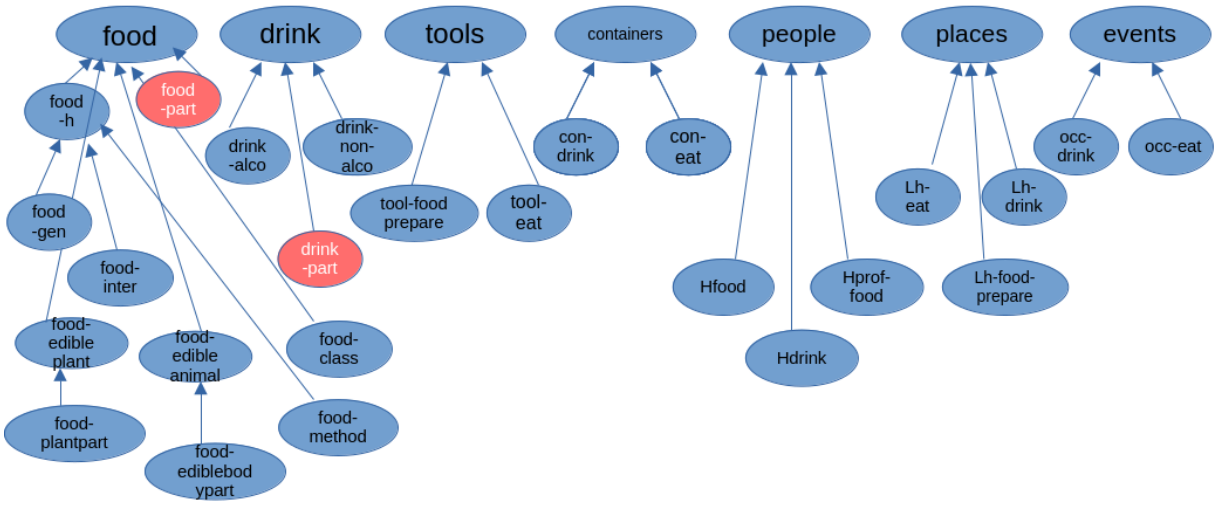


Figura 1: Categorias de comida e bebida

deles a comida e bebida, como *delícia* ou *acompanhamento*. Além disso, não podemos esquecer que os nomes dos animais e das plantas comestíveis se podem referir aos ditos animais e plantas sem ser num contexto de alimentação, e o mesmo se refere à palavra *água*, que forma rios e mares, e é usada por exemplo para lavar, e portanto não só para beber.[6]

Seja como for, os números das diferentes palavras que estão classificadas em cada categoria exigem uma clarificação: ainda não juntámos casos flagrantes de um mesmo conceito com várias grafias, quer devido a aportuguesamento de nomes estrangeiros (*uísque* e *whisky*, *chabli* e *Chablis*, *gim* e *gin*, *pudim* e *pudding*), como a variações ortográficas (*fêvera* e *febra*, *taverna* e *taberna*, *toicinho* e *toucinho*, *sande* e *sanduíche*, *água-ardente* e *aguardente*), algumas delas diferentes em Portugal e no Brasil. Também até agora considerámos casos de diminutivos ou aumentativos como potenciais palavras diferentes (*feijão-frade* e *feijão-fradinho*, *panela* e *panelão*), e aceitámos como unidades lexicais alguns casos de unidades multi-palavras, sobretudo quando tinham uma grafia alternativa com hífen, como *café da manhã* e *café-da-manhã*, *pé de moleque* e *pé-de-moleque*. Tudo isto terá de ser objeto de regras claras, numa próxima fase.

3 Desambiguação em contexto

Todas as palavras marcadas como potencialmente representando comida ou bebida têm de ser desambiguadas para se obter uma panorâmica deste campo semântico na literatura. Esse trabalho, em progresso, é feito usando regras automáticas de disambiguação escritas pelos autores, usando a filosofia de colaboração humano-máquina descrita em Santos e Mota (2010).

Aqui vamos apresentar a desambiguação da palavra *café*, palavra interessante não só por ser muito frequente, mas por ter muitas acepções que interessa distinguir, três das quais no campo semântico em que estamos interessados.

Vejam-se exemplos de usos distintos da palavra *café* em obras literárias:

planta *Para levar a efeito este pensamento – o da destruição da planta abençoada, servem-se do de cultivar com largueza o* **café** *no interior das províncias onde até o presente se cultivou largamente a cana.* (Franklin Távora, *O Matuto*, 1888)

bebida *Pois agora, colegas, disse o abade sorvendo o último gole de* **café***, o que está a calhar é um passeio à fazenda.* (Eça de Queirós, *O crime do Padre Amaro*, 1875)

parte de refeição *Rosa aparecia ao* **café***, exalando do seu sorriso, ...* (Eça de Queirós, *Os Maias*, 1888) *Entre o queijo e o* **café***, demonstrou-me Quincas Borba que o seu sistema era a destruição da dor.* (Machado de Assis, *Quincas Borba*, 1891)

loja *Dos* **cafés** *em geral, e de como são característicos da civilização de um país* (Almeida Garrett, *Viagens na minha terra*, 1846)

[6]Um estudo preliminar dessa palavra na literatura portuguesa e brasileira (Santos, 2024) apontou para apenas 7,7% casos de *água* como bebida.

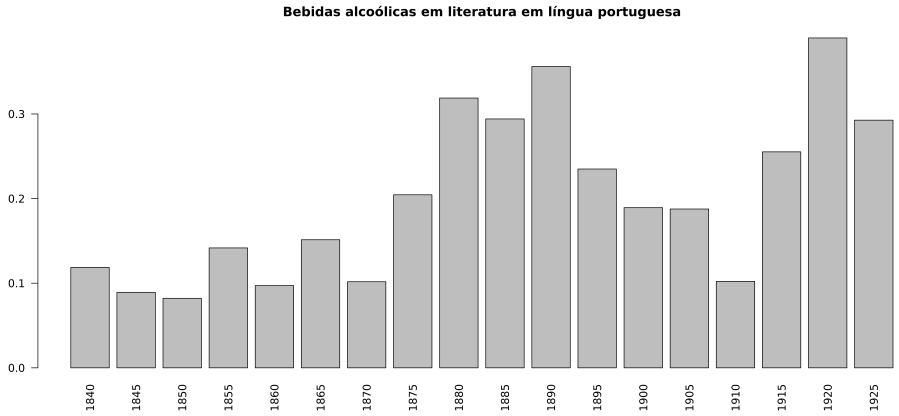


Figura 2: A frequência relativa das palavras referentes a bebidas alcoólicas, vezes 1000

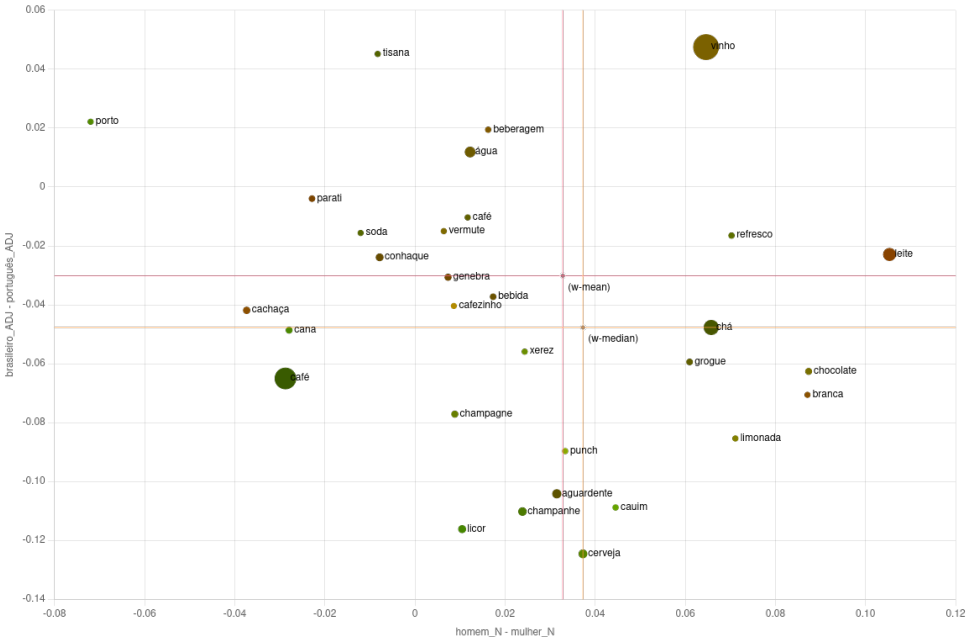


Figura 3: Bebidas, no eixo dos XX homens vs. mulheres, no eixo dos YY Brasil e Portugal

produto *Quem, ao vêr esta furia commercial que agita as populações, este ardor com que trocam lãs por* **café** *e assucar, algodões por cacáu e colza, cereaes por carvão de pedra, não se compadece dos destinos da humanidade* (António Pedro Lopes de Mendonça, *Memórias dum doido*, 1849)

Os casos de *café* presentes nas obras estudadas podem ser resumidos na Tabela 1.

Além disso, podemos observar diferenças entre a literatura portuguesa e brasileira, simplesmente dividindo as acepções entre as duas literaturas. Em primeiro lugar, o lema *café* aparece relativamente quatro vezes mais na literatura brasileira. Por outro lado, torna-se imediatamente claro que a referência aos cafés como loja é mais frequente em Portugal (apenas 13% na literatura brasileira) e que a referência ao café outros é mais frequente no Brasil (29%).

bebida	851	46%
lugar de comida	423	23%
refeição ou parte dela	163	8,8%
outros (planta, produto, etc.)	405	22%

Tabela 1: Classificação dos 1838 lemas *café*

3.1 Visualização dos dados

Usando o AC/DC (Santos e Bick, 2000) para fazer procuras complexas, e tratando o resultado com a ajuda da linguagem/ambiente R (R Development

Obra	Data	Autor	Palavras	Capítulos	Refeições	Comida	Comida relativa
Pupil	1867	JD	114426	42	3	211	1.84
Famil	1868	JD	147041	39	4	228	1.55
Morga	1868	JD	177122	33	8	204	1.16
Justi	1870	JD	34808	9	3	99	2.84
Fidal	1871	JD	168262	37	7	150	0.89
Padre	1875	EQ	171371	25	18	496	2.89
Filom	1884	AA	64569	23	3	112	1.73
Maias	1888	EQ	266017	18	21	857	3.22
Corti	1890	AA	95472	22	10	585	6.13
Sogra	1895	AA	61809	25	0	61	0.99
Total			1300897	273	77	30003	

Tabela 2: Informação sobre as dez obras analisadas, representando um total de 77 refeições em 273 capítulos. AA - Aluísio Azevedo, EQ - Eça de Queirós, JD - Júlio Dinis. Comida relativa foi multiplicada por 1000, ou seja, representa o número de palavras de comida por mil palavras.

Core Team, 2008), podemos ver a distribuição das bebidas alcoólicas por períodos de cinco anos na Figura 2.

Usando o CorpusEye (Bick, 2005), um ambiente especialmente desenvolvido para visualizar grandes conjuntos de dados anotados, podemos apreciar a relação entre bebidas, género do autor e contexto brasileiro ou português, usando palavras pulverizadas ("word embeddings"), na Figura 3. Embora este processo de visualização não distinga os diferentes sentidos das palavras, podemos ver que o vinho é preferencialmente português e o café e a cerveja brasileiros. Mais inesperadamente, o vinho é mais feminino e o café aparece em contextos mais masculinos.

4 Identificação de refeições

Uma refeição num romance é uma cena a partir da qual um leitor pode inferir período, classe social, às vezes até religião, com base no que e como as pessoas comem, e em que lugares. E pode servir para caracterizar protagonistas e relações entre eles, além de ilustrar tipos de convívio, e normas de etiqueta de uma época ou classe.

De acordo com Brown (1984, nossa tradução), "Refeições na ficção são acima de tudo signos literários: por isso, são sujeitos ao mesmo tipo de análise que outros fenómenos literários". McGee (2001) insiste na importância na ficção escrita por mulheres em inglês no princípio do século XX dos jantares ("dinners"), que, segundo ela, são a refeição mais social e aquela à qual estão associados mais rituais e expectativas. De acordo com esta autora, a reputação de uma mulher, na época, dependia geralmente das suas qualidades de cozinheira, ou, em estratos sociais mais elevados, da capacidade de organizar reuniões sociais à volta da comida.

Começámos por anotar manualmente as refeições (ou cenas à volta da comida) em dez romances em português, descritos na Tabela 2, para observar a relação entre o número de palavras do campo semânico comida e bebida e a existência de uma cena (refeição, preparação) envolvendo comida.

Na Figura 4, mostramos a densidade de palavras de comida e bebida por refeição por autor.

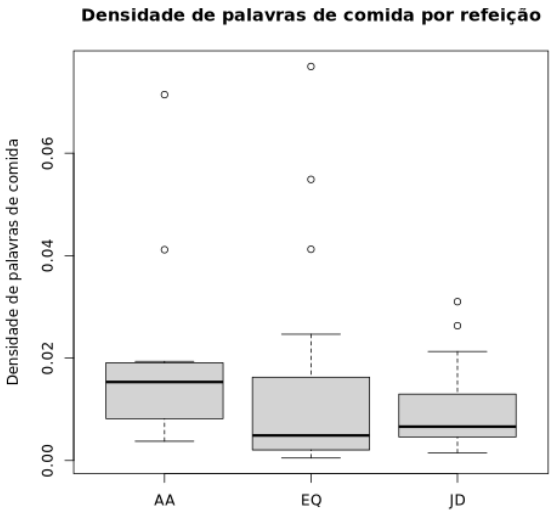


Figura 4: Quantas palavras do campo semântico comida e bebida por refeição, para cada autor

Nas Figuras 5 e 6, os gráficos do lado esquerdo mostram o tamanho das refeições na obra, por capítulo, enquanto que o os gráficos do lado direito

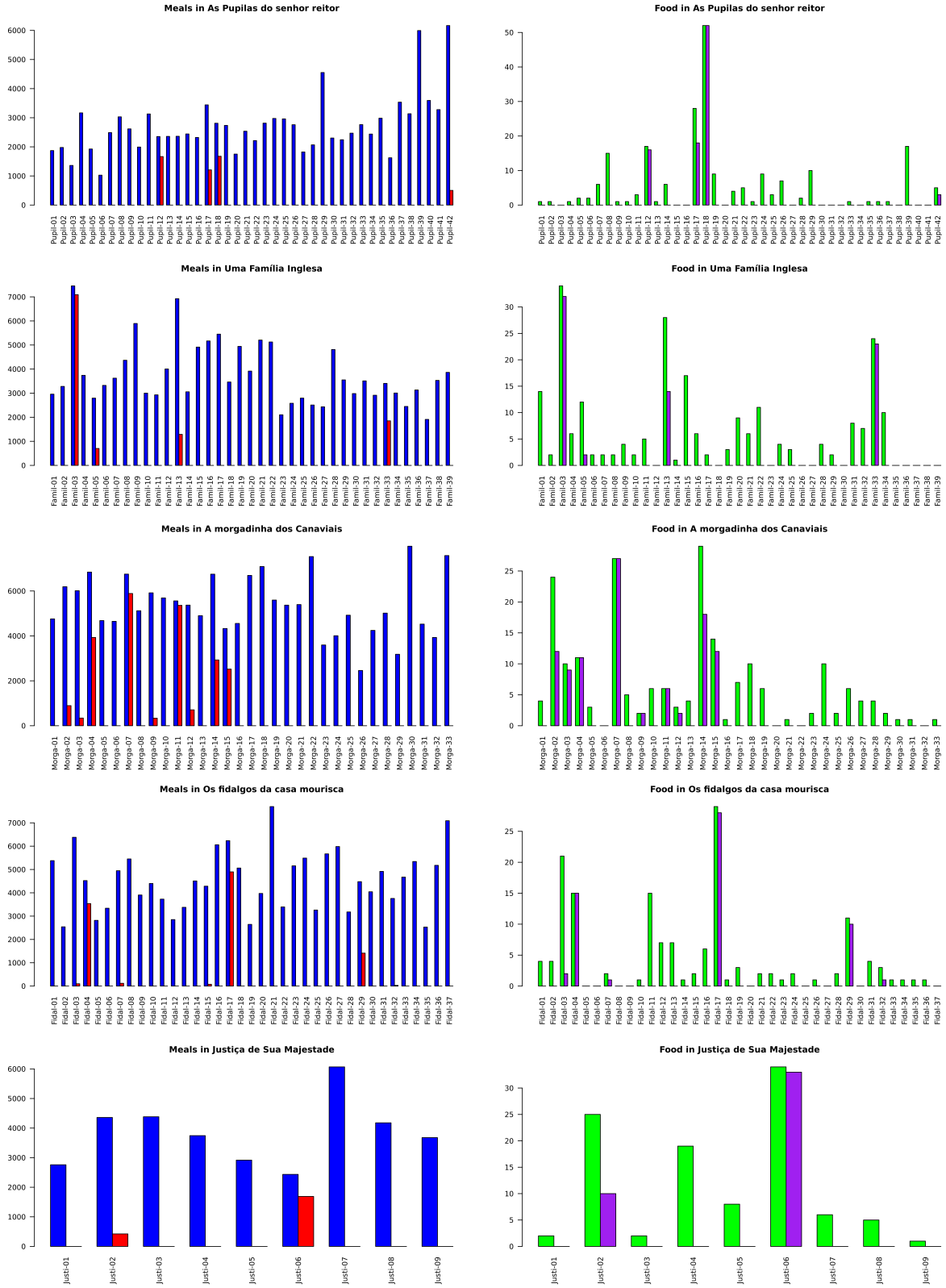


Figura 5: Tamanho das refeições em número de palavras, comparado com o tamanho dos capítulos, e palavras de comida por capítulo e por refeição, em cinco obras de Júlio Dinis

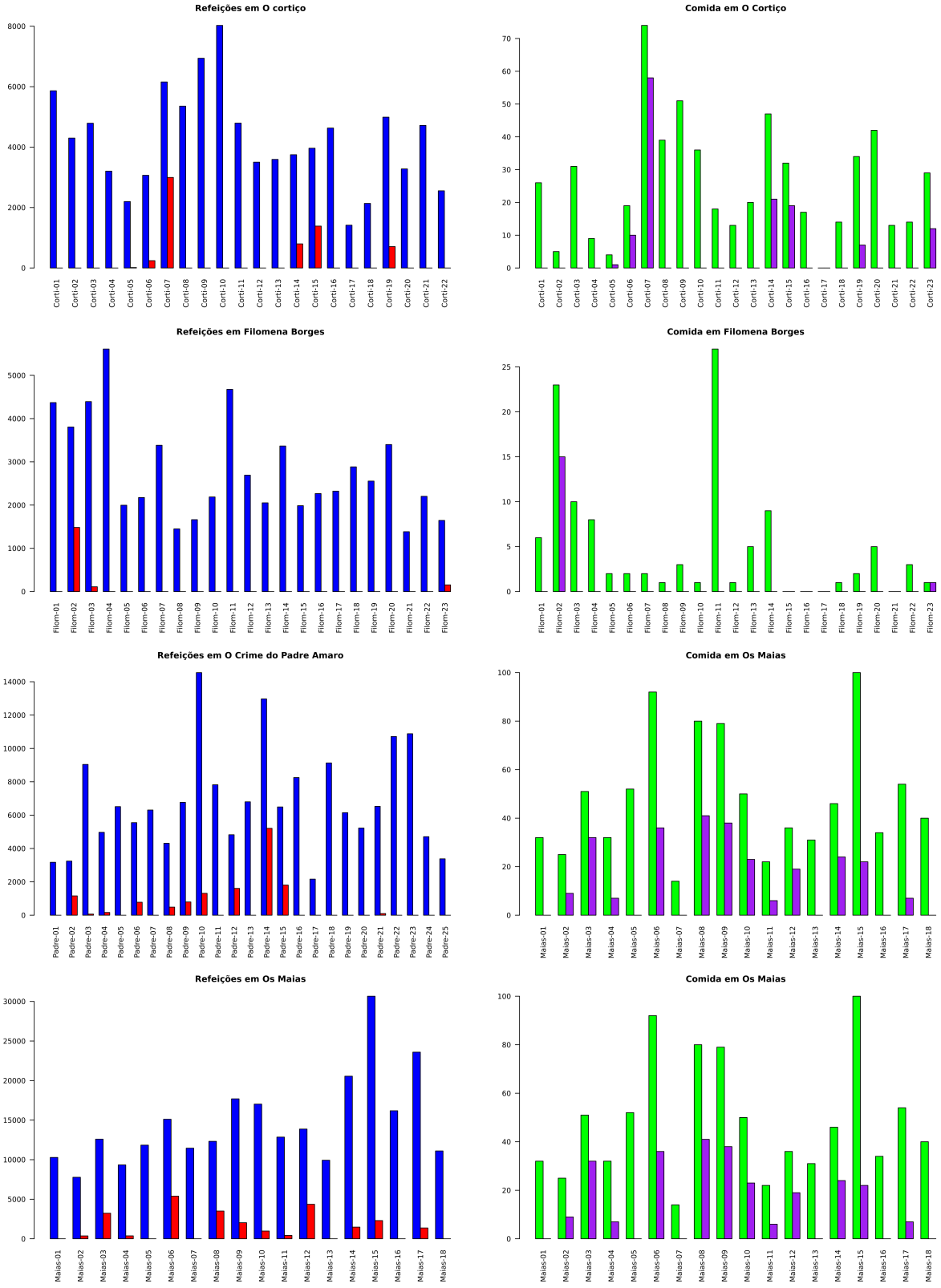


Figura 6: Tamanho das refeições em número de palavras, comaparado com o tamanho dos capítulos, e palavras de comida por capítulo e por refeição, em duas obras de Aluísio Azevedo and duas obras de Eça de Queirós

ilustram as palavras de comida e bebida por capítulo (a verde) e dentro de uma refeição (a roxo).

Estes dados mostram que uma densidade alta de palavras de comida e/ou bebida não é (não pode ser) o único critério para identificar uma refeição. Pelo contrário, até vemos que quanto maior (mais longa) é uma refeição literária (e, portanto, provavelmente mais importante para o enredo), menor é a densidade de palavras de comida nesse episódio. Além disso, vemos que existem muitas referências a palavras associadas ao campo de comida fora de refeições.[7] Pensamos explorar outras alternativas (ou em conjunção com a densidade) para identificar refeições literárias (Santos et al., 2026b):

- Procurar pistas sobre o início ou fim de uma refeição, por exemplo na forma de convite, ou indicação de que "o jantar está servido", ou de "levantar-se da mesa", como forma de sinalizar o começo ou fim de uma refeição. (Contudo, na maior parte das vezes estes sinais estão ausentes do texto, que muitas vezes "começa" já no meio de uma refeição.)
- Procurar pistas sobre sequências conhecidas. Uma refeição pode começar por um aperitivo ou terminar no café. Se isso for explicitado, podemos concluir que estamos dentro de uma refeição, e não uma simples menção de comida no texto.
- Proceder parágrafo a parágrafo, começando pelos parágrafos com alta densidade, e tentar juntá-los aos precedents ou seguintes se não houver mudanças radicais, por exemplo nos intervenientes ou no tempo.
- Usar os próprios nomes das refeições, se aparecerem na obra.

Outro elemento importante deste projeto é, além da identificação dos episódios associados a comida, a sua caracterização, para permitir posteriormente uma leitura distante.

Como tal, contamos associar, a cada episódio, várias características. Algumas objetivas, que reputamos serem possíveis de automatizar mais tarde, como

- número de participantes
- tempo do repasto/preparação
- nome da refeição (se mencionado)
- local (dentro ou fora de casa, num espaço privado ou público (café, taberna, restaurante...)
- tipo: preparação, refeição, brinde
- presença de criados (que servem)
- discurso direto durante a refeição

e outras, mais complexas mas provavelmente as mais relevantes de um ponto de vista literário, como

- objetivo (literário) da refeição no enredo (por exemplo, encontro de personagens, segredos revelados, altercação)
- objetivos do autor (caracterização das personagens, caracterização do local, do tempo, da classe social)

Em relação a este tipo de informação sobre as cenas, só depois de reunir um extenso grupo de casos anotados por peritos de literatura (e com suficiente consenso) poderemos avaliar se é possível obter uma semi-automatização desta tarefa,

5 Observações finais

Pensamos que este recurso – corpos anotados e desambiguados sobre o tema da comida e bebida – pode servir para estudar essa parte da cultura. Porque como diz Eagleton (1997, página 25) em *Edible écriture*, "se há algo garantido sobre comida, é que não é nunca apenas comida"[8].

Este recurso será útil não só em termos lexicais, mas também para detetar metáforas associadas, para compreender o uso das palavras de comida e bebida em texto literário: caracterização económica, psicológica, histórico-social, assim como para a identificação de episódios no enredo (refeições ou cenas de preparação de comida), e qual o seu objetivo.

Neste artigo, descrevemos a identificação manual de refeições em textos literários em português no espírito da leitura distante e propusemos a sua automatização através de várias estratégias diferentes, nomeadamente: concentração de palavras referentes ao domínio da alimentação; identificação de sequências naturais numa refeição; e referência à própria refeição, além de debatermos como caracterizar as próprias refeições em texto literário.

[7] Embora tenhamos de reconhecer que estes dados ainda não são baseados numa completa desambiguação, mas sim de apenas 70% dos casos.

[8] "If there is one sure thing about food, it is that it is never just food".

Referências

Leonard Barkan. 2021. *The Hungry Eye: Eating, Drinking and European Culture from Rome to the Renaissance*. Princeton University Press.

Eckhard Bick. 2005. CorpusEye: Et brugervenligt web-interface for grammatisk opmærkede korpora. Em *10. Møde om Udforskningen af Dansk Sprog 7.-8.okt.2004, Proceedings*, páginas 46–57. Århus University.

Charlotte Boyce e Joan Fitzpatrick. 2017. *A History of Food in Literature From the Fourteenth Century to the Present*. Routlegde.

James W. Brown. 1984. *Fictional Meals and their Function in the French novel: 1789-1848*. Univ. of Toronto Press.

J Michelle Coghlan. 2020. *The Literature of Food*. Cambridge University Press.

Terry Eagleton. 1997. Edible écriture. The Times, Oct. 24, 1997.

Michiel Korthals. 2008. Food as a Source and Target of Metaphors: Inclusion and Exclusion of Foodstuffs and Persons through Metaphors. *Configurations*, 16:77–92.

Diane McGee. 2001. *Introduction: A Time to Eat. In Writing the Meal: Dinner in the Fiction of Twentieth-Century Women Writers*. University of Toronto Press.

Anna Meigs. 1997. Food as cultural construction. Em *Food and culture: a Reader*, páginas 95–106. Routledge.

Franco Moretti. 2013. *Distant Reading*. Verso.

R Development Core Team. 2008. R: A language and environment for statistical computing.

Diana Santos. 2014. *Corpora at Linguateca: Vision and Roads Taken*, páginas 219–236. Bloomsbury.

Diana Santos. 2019. Literature studies in literateca: between digital humanities and corpus linguistics. Em *Humanists and the digital toolbox: In honour of Christian-Emil Smith Ore*, páginas 89–109. Novus forlag.

Diana Santos. 2024. Distant reading of food and drink in three languages.

Diana Santos e Eckhard Bick. 2000. Providing Internet access to Portuguese corpora: the AC/DC project. Em *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC 2000)*, páginas 205–210.

Diana Santos e Cláudia Freitas. 2019. Estudando personagens na literatura lusófona. Em *STIL - Symposium in Information and Human Language Technology*.

Diana Santos, Elizaveta Khachaturyan, Michael Preminger, Åse Kristine Tveit, e Eckhard Bick. Presenting Reading Food and its infrastructure. *Digital Humanities in the Nordic and Baltic Countries Publications*.

Diana Santos, Luisa Lima, e Emanoel Pires. 2026a. Marcação de correferência para a caracterização de personagens em obras literárias em português. Em *Proceedings of PROPOR 2026*.

Diana Santos e Cristina Mota. 2010. Experiments in human-computer cooperation for the semantic annotation of Portuguese corpora. Em *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2010)*, páginas 1437–1444. European Language Resources Association.

Diana Santos, Michael Preminger, Åse Kristine Tveit, e Elizaveta Khachaturyan. 2026b. Identifying food episodes in literary texts. In preparation.

Roger C. Schank e Robert P. Abelson. 1977. *Scripts, Plans, Goals, and Understanding: An Inquiry into Human Knowledge Structures*. Routledge.

Christof Schöch, Tomaz Erjavec, Roxana Patras, e Diana Santos. 2021. Creating the European Literary Text Collection (ELTeC): Challenges and Perspectives. *Modern Languages Open*, 1:1–19.

Véronique Traverso e Loubna Dimachki. 2017. "ktîr Tajjibe ce plat! fî garlic?": Compliments and assessments in French and Lebanese dinner talk. *Intercultural Pragmatics*, 4(2):137–163.

Fauna e Flora setecentista: das entidades aos problemas de normalização

Helena Freire Cameron[1,2], **Fernanda Olival**[2], **Daniel Reyes**[2], **Renata Vieira**[2]
[1]Portalegre Polytechnic University, Portugal
[2]University of Évora, CIDEHUS
helenac@ipportalegre.pt, mfo@uevora.pt
daniel.a.g.reyes@gmail.com, renatav@uevora.pt

Resumo

Este artigo aborda tarefas do tratamento de fontes históricas do século XVIII, em língua portuguesa. O trabalho desenvolvido incidiu nos domínios específicos de fauna e flora. Por esta última característica, esperava-se um fraco nível de ambiguidade vocabular, mas assim não aconteceu. Por isso, apresenta-se um roteiro do processo de normalização ortográfica; descreve-se a constituição do *corpus* anotado de entidades e, sobretudo, discutem-se problemas ligados à variação lexical nestes *thesauri* de especialidade e os constrangimentos do processo. Desta forma, pretende-se contribuir para a reflexão sobre o que é o processo de normalização de fontes históricas e chamar a atenção para a importância das boas práticas neste quadro.

1 Introdução

A anotação de entidades tem possibilitado estudos de diversa natureza em variadas áreas. No entanto, para uma realidade pretérita, a anotação de EN reveste-se de uma complexidade acrescida, com exigência de uma definição mais diferenciada ao nível das categorias, uma vez que as comummente aplicadas não são suficientemente abrangentes para o universo específico aqui tratado: Fauna e Flora.

No campo das Humanidades tem-se insistido quase sempre no mesmo padrão de categorias e num leque reduzido de entradas (Rodríguez-Puente et al., 2019). Acresce que a anotação com vista à criação de *datasets* capazes de constituir padrão ouro para treino de modelos tem requisitos que devem ser cumpridos, especialmente em textos históricos, pois exigem categorias adaptadas ao domínio específico e às linhas estruturantes da época em estudo (Álvarez Mellado et al., 2021).

Neste trabalho, aborda-se uma fonte histórica do século XVIII que reúne dados relevantes sobre o território português, a sua ocupação e o seu enquadramento natural (serras, rios) nesse período. Duas dimensões diretamente associadas à ocupação do território são a fauna e a flora, frequentemente relacionadas não apenas com os hábitos e costumes das populações, mas também com as atividades económicas das diferentes regiões.

A constituição de *corpora* anotados a partir de textos históricos em português e com validação científica humana é ainda reduzida face a outras línguas igualmente de expressão mundial. Citem-se os trabalhos desenvolvidos por Aguilar et al. (2017); Grilo et al. (2020); Zilio et al. (2022); Santos et al. (2024); Nunes et al. (2025) entre outros. Estes estudos e recursos são muito necessários, não só como aplicação de processos de Processamento de Linguagem Natural (NLP) a textos pré-contemporâneos como pela constituição de *datasets* capazes de se constituírem como *gold standard*, por exemplo, para tarefas automatizadas de anotação de entidades.

Neste descreve-se a constituição de um *corpus* anotado em duas áreas específicas, Flora e Fauna. Analisam-se os dados relativos às subcategorias e às entidades, em cinco concelhos do Sul de Portugal no século XVIII, que serviram de amostra. Com base nos resultados obtidos da anotação, discutem-se tópicos que vão muito para além da normalização gráfica, nomeadamente questões de variação lexical e constrangimentos neste *corpus* anotado que, sendo em domínios específicos, se esperava, à partida, que fosse mais objetivo e isento de ambiguidades.

2 A Constituição do *Corpus* anotado

As *Memórias Paroquiais* são um conjunto textual de grande relevância, reunindo informações de cada uma das freguesias do Portugal de setecentos, após o sismo de 1755, que teve impacto sobre boa parte do território. A coroa portuguesa fez chegar a todos os párocos um questionário com 60 perguntas sobre a Terra (população, edificado, usos e costumes, etc.), a Serra (características do território,

Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)
Vol. 2, pages 275–281

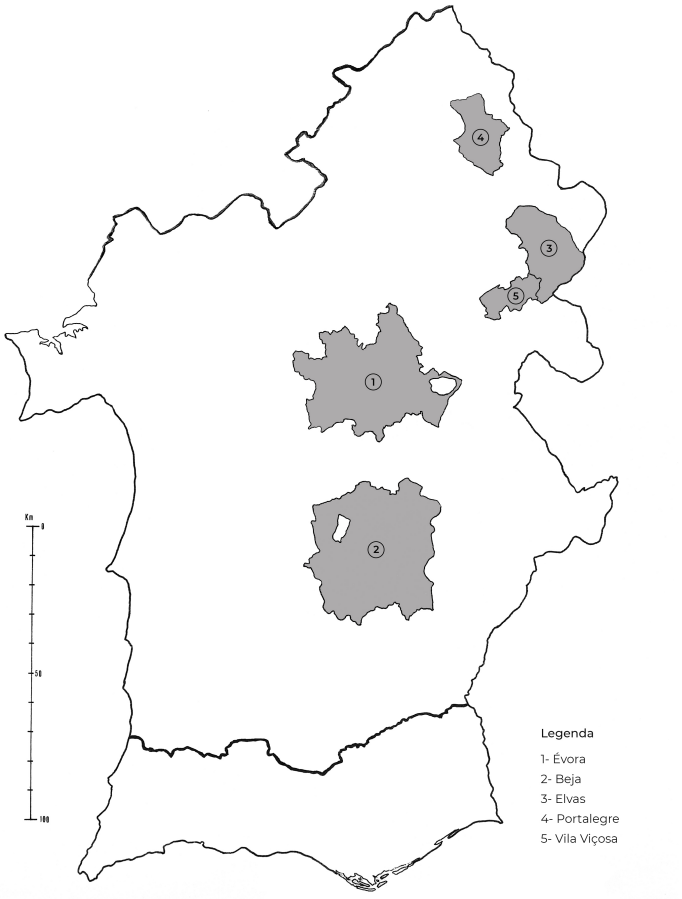


Figura 1: Região sul de Portugal, com as áreas dos concelhos estudados

plantas e animais, entre outros) e os Rios (existências, navegabilidade, rentabilidade económica, propriedades das águas, entre outros). As respostas dos padres, que visavam ser reunidas num futuro Dicionário Geográfico, foram coligidas posteriormente. As digitalizações dos originais manuscritos estão disponíveis *online* no Arquivo Nacional da Torre do Tombo. Os textos relativos ao Sul de Portugal foram transcritos por paleógrafos e estão disponíveis no repositório do CIDEHUSDigital[1]. Foi a partir destes que se desenvolveu esta investigação.

Os textos históricos anteriores ao século XX caracterizam-se por uma grafia não padronizada. Uma palavra, por pequena que fosse, podia ser redigida de múltiplas formas, por vezes no mesmo documento. Foi apenas na segunda década do séc. XX que foi introduzida a ortografia uniforme que todos deviam seguir. Por conseguinte, para recuperar informação de forma eficaz, a realidade anterior representa hoje um grande constrangimento. Assim, uma tarefa essencial é tentar normalizar a grafia, seja de modo manual ou tentando automatizar. Este último desiderato é um objetivo ainda em construção, conforme Cameron et al. (2023).

Para este estudo, os textos foram normalizados manualmente para a ortografia contemporânea, preservando toda a variação lexical. Para não desvirtuar a realidade histórica e linguística, optou-se por uma intervenção conservadora, limitada à atualização gráfica para o padrão hodierno: regularizaram-se os ditongos nasais (am → ão), eliminaram-se consoantes pseudoetimológicas (e.g. -th-) e reduziram-se consoantes duplas não etimológicas (e.g. -ll-, -bb-). Manteve-se a variação linguística (e.g. oiro/ouro), bem como formas antigas ainda em uso (El-Rei, cousa, mui). Ainda assim, o processo revelou-se complexo: na ausência de automatização para a normalização gráfica, todo o trabalho foi feito manualmente, o que implicou um elevado investimento de tempo e de recursos humanos.

Concluída esta tarefa, foi feita uma ano-

[1]https://www.cidehusdigital.uevora.pt/

tação manual de entidades de Fauna e Flora na plataforma INCEpTION[2]. Os dados obtidos (formato CONLL) foram pós-processados, tendo sido lematizados manualmente (por entidade), constituindo assim o dataset. A anotação manual nos domínios de Fauna e da Flora foi feita em 87 textos relativos às três capitais de distrito do Alentejo (Évora, Beja e Portalegre), na região do sul, que correspondia a cerca de um terço de Portugal. Às localidades invocadas juntaram-se mais dois concelhos: Vila Viçosa, por ter sido sede da Casa de Bragança até 1640, e Elvas, pela sua posição fronteiriça. Na Figura 1, pode observar-se a localização geográfica de todos os concelhos tratados.

Évora era o concelho com maior relevância à época. Era composto por 22 freguesias, fazendo parte de uma ampla zona rural. Évora constituía, à data, uma urbe com importância económica e político-administrativa: era sede de arcebispado e de um dos três tribunais do Santo Ofício do espaço metropolitano português; tinha universidade. Até 1640 fora a segunda cidade portuguesa, em termos políticos.

Beja, no Baixo Alentejo, era uma zona predominantemente rural, com propriedades de grande extensão. Contabilizava, em 1758, 28 freguesias, urbanas e rurais.

Elvas situava-se no Alto Alentejo e reunia, à época, 17 freguesias.

O que é hoje o concelho de Portalegre congregava, no século XVIII, 10 freguesias, urbanas e rurais. Este concelho apresenta uma geografia muito distinta dos restantes por se situar numa zona fortemente arborizada, em plena Serra de S. Mamede, no Alto Alentejo.

Vila Viçosa englobava, na época, 6 freguesias, urbanas e rurais. Situa-se no Alentejo Central e geograficamente contém várias serras, sendo cruzada por um rio, afluente do Guadiana.

3 Fauna, Flora e sub-categorias

As categorias Fauna e Flora, pela sua abrangência, foram fracionadas em unidades mais pequenas, que pudessem descrever melhor as várias tipologias dentro de cada uma destas. Assim, **Fauna** foi subdividida em sete subcategorias (em inglês, para maior comparabilidade): *Fish* (peixes), *Bird* (aves), *Mammal* (mamíferos), *Reptile* (répteis), *Insect* (insetos), *Other* (outros animais não incluídos nos itens anteriores), *Product* (produtos derivados, como couro, etc.). No que respeita a **Flora**, esta foi igualmente dividida em sete subcategorias: *Herb* (ervas silvestres e cultivadas), *Tree* (árvores), *Cereal* (cereais), *Vegetable* (verduras e legumes), *Fruit* (frutas), *Other* (outros elementos de Flora não incluídos nos itens anteriores), *Product* (produtos derivados ou transformados, como cortiça, vinho, azeite, etc.)

Nos oitenta e sete textos foram anotadas 1068 ocorrências. Estas foram lematizadas e correspondem a 208 entidades distintas. Este procedimento de lematização é também uma tarefa fundamental, de modo a obter dados de ocorrências mais fiáveis, permitindo anular as flexões em género e número, neste caso.

Em **Fauna**, anotaram-se 53 entidades, que pertencem a três subcategorias: Bird, Mammal, Fish e *Product*. Não se encontraram nos textos em apreço outras. A subcategoria Mammal descreve mamíferos, produzidos tanto para consumo de carne/ leite (ovinos, caprinos, bovinos, suínos e leporídeos). como outros associados à pastorícia, como javali, lebre e veado; inclui também animais selvagens, como lobo, raposa e gineto/gato bravo. A subcategoria Fish contém 20 entidades, constituídas por peixes de rio.

No que respeita à **Flora**, foram anotadas 174 entidades, nas subcategorias Cereal, Fruit, Herb, Tree, Vegetable, Products e Other. A subcategoria que tem maior número de entidades é Herb (64), reunindo um conjunto de várias espécies, quer medicinais, quer o que hoje se chamam aromáticas, ou mesmo plantas selvagens. Este agregado fornece uma importante "radiografia" não só da existência e cultivo das espécies vegetais como dos usos que se faziam das plantas, por exemplo o tratamento de enfermidades (e.g. erva da erisipela), a sua direta implicação em atividades económicas (carrasco, que servia para a produção de tintas), práticas supersticiosas (e.g. arruda), etc. No que diz respeito à subcategoria Fruit, foram anotadas 43 entidades, entre frutos frescos para a alimentação (melão, ameixa, romã, melancia, maçã, etc.), frutos secos (passas de figo, amêndoa, noz) e leguminosas (feijão branco, feijão frade).

4 Análise dos Dados

Neste itinerário de trabalho, com tarefas sequenciais bem delimitadas, conhecer os dados é igualmente importante para definir domínios de atuação, tendo em vista normalizar com mais eficácia.

[2]https://inception-project.github.io/

Os dados do *corpus* anotado foram ordenados por subcategorias, de modo a poder ver-se quais as dominâncias e/ou casos isolados.

Na categoria **Fauna** só se encontraram três subcategorias: *Mammal*, *Fish* e *Bird*, conforme a tabela 1.

Categoria	Ocorrências	Entidades distintas
FAUN_MAMMAL	131	31
FAUN_FISH	129	20
FAUN_BIRD	21	2

Tabela 1: Distribuição de dados sobre Fauna

Na categoria **Flora**, ao invés, foram encontradas todas as sete subcategorias, hierarquizadas na tabela 2:

categoria	ocorrências	entidades distintas
FLORA_CEREAL	362	6
FLORA_FRUIT	110	43
FLORA_HERB	86	64
FLORA_TREE	71	25
FLORA_VEGET	66	18
FLORA_PROD	55	7
FLORA_OTHER	34	11

Tabela 2: Distribuição de dados sobre Flora

No que respeita às entidades, estas foram ordenadas por frequência decrescente, de que a Figura 2 é elucidativa:

Figura 2: Word cloud da totalidade de entidades anotadas

A subcategoria *Cereal* tem apenas 6 entidades distintas, mas é aquela que regista maior número de ocorrências (362). A segunda subcategoria mais frequente é *Mammal*, com 131 ocorrências e 31 entidades distintas. A terceira mais frequente é *Fish*, com 129 ocorrências. A elevada frequência destas entidades é reveladora da importância das atividades piscícolas nas bacias hidrográficas da região, tendo sido anotadas 20 entidades distintas. As denominações dos peixes foram uma das vertentes deste estudo que causaram maiores constrangimentos, conforme se discutirá adiante.

Fruit é a quarta subcategoria mais frequente, tendo sido anotadas 43 entidades distintas, que bem demonstram a variedade de espécimes existentes nestes concelhos.

A quinta subcategoria mais recorrente é *Herb*. Esta é a que contém maior número de entidades distintas e foi onde também o processo de normalização não foi fácil, como se discutirá mais à frente. A subcategoria com menor número de ocorrências é **Flora** *Other*.

Vejam-se agora as subcategorias em relação aos concelhos onde foram anotadas.

Em 1758, quase todas as freguesias ou paróquias do Alentejo teriam uma parte urbana e outra rural. A localização geográfica e características naturais influenciaram, certamente, o maior ou menor registo de entidades nestes domínios específicos da **Fauna** e **Flora**.

Observando ao nível dos concelhos quais as subcategorias mais frequentes verifica-se que, nos concelhos de Évora, Beja, Portalegre e Elvas, a subcategoria mais frequente é *Cereal*. A única exceção era constituída por Vila Viçosa, onde a subcategoria mais referida foi *Herb*, conforme Figura 3.

Inequivocamente, trigo é o cereal mais referido nestes textos, revelando a sua importância nestas regiões. Os cereais anotados, com o respetivo número de observações, resumem-se a centeio (55), cevada (99), pão (12), tremês (2) e trigo (194). Note-se que, na época, "pão" era equivalente a "cereal".

Ainda referente aos espécimes vegetais, os frutos são uma subcategoria com elevada frequência e grande número de entidades. Anotam-se diversos frutos para consumo humano. Destaca-se, ainda, a expressão "pêros de guarda", que se reportaria a uma espécie de maçã de maior durabilidade.

Os produtos derivados evidenciam a atividade económica destas regiões do sul de Portugal, algumas ainda hoje subsistentes. Em 1758, o azeite era o recurso com maior número de ocorrências (29), seguindo-se o vinho (16), a farinha (4) e a tinta (1).

As entidades da subcategoria **Flora** *Tree* têm elevadas frequências nos concelhos de Évora e Portalegre, e baixas frequências nos restantes concelhos. Englobam 26 entidades distintas. São sobretudo árvores com aproveitamento económico, como sobreiro, oliveira e azinho, ainda hoje de grande importância.

Os frutos surgem com elevada frequência em Elvas, contrastando com a baixa incidência em Évora e Beja. Ainda no presente, a “ameixa de Elvas”

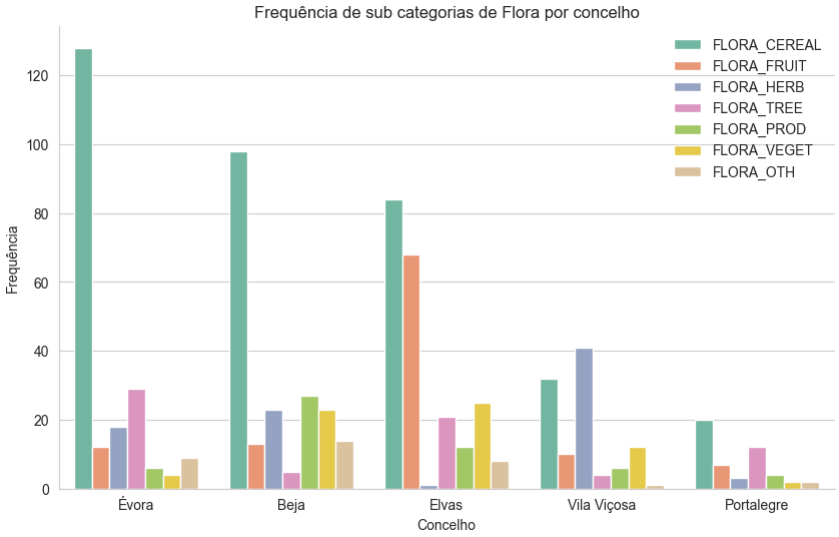


Figura 3: Subcategorias de Flora por concelho

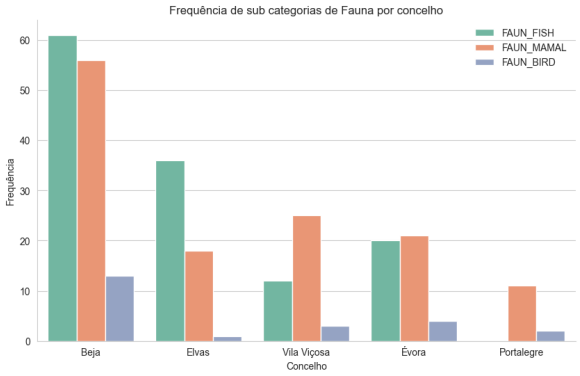


Figura 4: Subcategorias de Fauna por concelho

mantém-se relevante, afirmando-se como um produto de reconhecido valor económico.

O "vinho", atualmente com grande impacto na economia de todo o Alentejo, tinha um modesto número de registos no que foi o município de Portalegre em meados do século XVIII.

Relativamente aos espécimes animais, destacam-se os animais de criação, como “cabra”, “coelho” e “porco”, este último ainda hoje com elevado valor económico. Registam-se também animais selvagens, como a “perdiz”, espécie cinegética que continua a ser muito procurada. Entre os peixes de rio mais referidos, salientam-se o “bordalo” e a “boga”. Todos os concelhos têm anotações de Fauna, embora com diferente número de ocorrências, como se pode ver na Figura 4.

Relativamente às entidades da subcategoria Fish, estas têm maior número de ocorrências nos concelhos de Évora, Beja e Elvas. Em Vila Viçosa tem um número de ocorrências reduzido face aos concelhos anteriores; em Portalegre, é mesmo inexistente, o que é inverosímil e dever-se-á à falta de atenção dos párocos que responderam ao inquérito.

5 A variação lexical e os seus problemas

Em domínios de especialidade, como denominações geográficas, nomes vulgares de plantas e animais, há grande discrepância entre as formas do século XVIII e os equivalentes atuais para efeitos de normalização.

No *corpus* anotado que serviu de base para este estudo, pode observar-se que há uma relação direta entre maior número de entidades diferentes numa subcategoria e maiores constrangimentos em sede de normalização. As baixas frequências de vocabulário foram as que trouxeram maiores dificuldades. Estes constrangimentos podem ser devidos a vários fatores. O primeiro pode ser assumido como provável erro, ou do escrevente, ou do transcritor, obrigando a uma consulta do original manuscrito para verificação do termo.

O segundo constrangimento tem a ver com as próprias denominações. Os padres respondentes não seriam, certamente, botânicos, nem biólogos, pelo que as denominações de plantas e peixes nas *Memórias Paroquiais* poderão ser designações regionais, sem correspondência na atualidade.

A consulta a especialistas trouxe, igualmente, grandes desafios. Perante a inexistência de imagens quer das plantas quer dos animais, uma vez que a fonte é apenas textual, a classificação taxonómica dos espécimes torna-se mais complexa. Veja-se o exemplo de duas denominações de peixes: "combo

beiçudo" e "cabecinha". Estas poderão ser espécies de barbos, peixe muito frequente na bacia do rio Guadiana, mas estas denominações não existem na atualidade. Igualmente, algumas denominações podem referir-se ao estado juvenil destes peixes e não propriamente a uma subespécie.

No que respeita à Flora, algumas espécies de plantas não cultivadas carecem igualmente de verificação por especialistas. Um exemplo é a multiplicidade de entidades designadas por cardos, como cardo abrolho, cardo alvarinho, cardo arzol, cardo corredor, cardo rasteiro, cuja denominação atual (nome vulgar) tem de ser validada, uma vez que a correspondência destas denominações para a atualidade não é imediata.

Foi encontrado um registo de uma planta, "saisso", cuja regularização ortográfica não conseguiu ser validada para um possível registo ortográfico atual, já que esta denominação, tal como está, não existe em dicionários nem é conhecida pelos especialistas consultados.

Uma outra dificuldade, que é paralela à normalização gráfica mas que, em sede de enriquecimento de dados, tem de ser acautelada, tem a ver com variantes lexicais que designam a mesma espécie. Nos textos, no que respeita a animais, encontramos chibo/cabra, carneiro/borrego, gineto/gato bravo, raposa/zorra, designações que, no par, são consideradas equivalentes. Contudo, também aqui pode haver um uso regional ou uso da forma mais "culta" em detrimento da popular, por exemplo em chibo/cabrito. Para as espécies vegetais, são usados como variantes: bolota/lande, pão/cereal, azinho/ azinho-sobro/ azinho-sôvero, maçã/pêro, designações que, igualmente, podem admitir um uso regional.

Pensando-se que, sendo esta uma área de especialidade, o vocabulário seria mais limitado e mais preciso, mas verificou-se exatamente o contrário. Os padres, não sendo especialistas em botânica e zoologia, usariam termos comuns. Por outro lado, o registo escrito poderá estar ligado à oralidade pelo que poderá conter erros, ou usos regionais. Assim, em domínios de especialidade, a normalização (orto)gráfica não consegue ser apenas uma simples atualização da grafia requerendo uma intervenção de especialistas de domínio para confirmação do registo escrito normalizado a adotar, que não alterará nunca a variação lexical. No processo aqui descrito, esta verificação foi realizada na nomenclatura de dicionários autorizados de língua portuguesa e por especialistas nos domínios da Fauna e da Flora ou da Geografia. Ainda assim, algumas denominações, sobretudo de peixes e de plantas silvestres, não conseguiram reunir, nesta fase, a unanimidade dos especialistas, constituindo isto uma limitação ao trabalho desenvolvido.

Outro constrangimento teve a ver com a validação possível das espécies para uma futura constituição de datasets enriquecidos. Os especialistas tiveram acesso a dados apenas textuais, sem existência de nomes científicos, o que dificultou muito o trabalho de validação científica das denominações. Acresce que, atendendo a que apenas se analisaram freguesias de cinco concelhos, a existência de designações locais face a outras de outras regiões não conseguiu ser completamente comprovada neste estudo, constituindo igualmente uma limitação. Todavia, ficou o alerta. Em futuros trabalhos, com maior número de freguesias, estas entidades serão revalidadas. Não está posto de parte o recurso a trabalho colaborativo, aberto a quem conhece a realidade local, por vezes de escala micro, para ajudar a identificar estes recursos ou micro-topónimos; pode ser uma alternativa, embora algumas espécies possam já ter desaparecido.

6 Conclusão

Como se demonstrou, pensar em normalização automática de textos históricos, com grande abrangência temática, incluindo vocabulário de setores específicos, representa um grande desafio. Implica estar aberto a integrar especialistas em domínios que se situam muito para além das Humanidades. Isto torna-se bem evidente quando se pensa em abarcar especialidades como Fauna e Flora.

Normalizar está longe de ser uma tarefa de resultados imediatos. Exige ampla preparação para ser consistente. Normalizar textos históricos é mais do que um mero ajuste ortográfico. É um processo que exige um trabalho interdisciplinar prévio para desambiguar com segurança. Reveste-se de uma importância crucial para uma melhor compreensão dos próprios textos e da realidade histórica em questão. Para além disso, permite trazer para o presente textos de elevado valor patrimonial que, por via, passam a poder ser lidos por públicos generalistas e de especialidade.

Agradecimentos

Este trabalho é financiado por fundos nacionais através da Fundação para a Ciência e Tecnologia (FCT - Portugal), no âmbito do projeto

UID/00057/2025.

References

Gustavo Aguilar, Suraj Maharjan, Adrian Pastor López Monroy, and Thamar Solorio. 2017. A multi-task approach for named entity recognition in social media data. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 148–153.

Helena Freire Cameron, Fernanda Olival, and Renata Vieira. 2023. Planear a normalização automática: tipologia de variação gráfica do corpus das memórias paroquiais (1758). *LaborHistorico*, 9 (1):2359–6910.

Sara Grilo, Márcia Bolrinha, João Silva, Rui Vaz, and António Branco. 2020. The BDCamões collection of Portuguese literary documents: a research resource for digital humanities and language technology. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 849–854, Marseille, France. European Language Resources Association.

Rafael Oleques Nunes, Joaquim Santos, André Spritzer, Dennis Giovani Balreira, Carla Dal Sasso Freitas, Fernanda Olival, Helena Freire Cameron, and Renata Vieira. 2025. *Assessing European and Brazilian Portuguese LLMs for NER in Specialised Domains*, volume 15412. Springer, Cham.

Paula Rodríguez-Puente, Cristina Blanco-García, and Iván Tamaredo. 2019. Annotation in the corpus of historical english law reports (chelar): Potential for historical genre analysis. *Journal of the Spanish Association for Anglo-American Studies*, 41 (2):63–84.

Joaquim Santos, Helena Freire Cameron, Fernanda Olival, Fátima Farrica, and Renata Vieira. 2024. Named entity recognition specialised for portuguese 18th-century history research. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese*, pages 117–126.

Leonardo Zilio, Maria Jose Bocorny Finatto, and Renata Vieira. 2022. Named entity recognition applied to Portuguese texts from the 18th century. In *Proceedings of the Second Workshop on Digital Humanities and Natural Language Processing (2nd DHandNLP) co-located with International Conference on the Computational Processing of Portuguese (PROPOR 2022) Virtual Event, Fortaleza, Brazil, CEUR Workshop Proceedings, v. 3128*.

Elena Álvarez Mellado, María Luisa Díez-Plata, Pablo Ruiz-Fabo, Helena Bermúdez, Salvador Ros, and Elena González-Blanco. 2021. Tei-friendly annonation scheme for medieval named entities: a case on a spanush medieval corpus. *Language Ressources and Evaluation*, 55 (2):525–549.

Exploring automatic terminology extraction from historical medical data

Leonardo Zilio
CENTAL, UCLouvain, Belgium
leonardo.zilio@uclouvain.be

Maria José Bocorny Finatto
PPGLetras, UFRGS, Brazil
mariafinatto@gmail.com

Abstract

This paper analyzes the performance of several terminology extraction methods when confronted with historical specialized texts that do not conform with modern orthographical norms. We tested two extraction methods based on linguistic patterns, four prompt-based generative artificial intelligence (GenAI) models, and one BERT-like model. Some of these models went through fine-tuning for terminology extraction, and one of these is specialized in the extraction of medical terms from documents written in Portuguese. For the GenAI models, we tested four different prompting strategies. As test set, we used chapter fifteen of the second part of the book *Aviso a' Gente do Mar sobre a sua Saude* [Advice to Sea People about their Health], originally written in French by G. Mauran at the end of the 18th century, and translated and adapted to Portuguese in 1794. The chapter was manually annotated with terminology, and the evaluation was conducted separately as an f-measure automatic evaluation, as well as a manual precision-based evaluation. This second evaluation method was applied to observe if the automatic extraction methods could complement the original token-based annotation. Results show that using automatic extraction methods to complement the manual annotation can improve coverage, even if individual models do not achieve high extraction quality. By combining two or more models though, a recall of more than 90% could be achieved in the test data.

1 Introduction

Automatic terminology extraction (ATE) is an important Natural Language Processing (NLP) task that serves as basis for several downstream linguistic and computational tasks, such as the lexicometrical analysis and consistent translation of specialized texts. ATE can also further advance research in Digital Humanities, as it contributes to the description and understanding of historical practices in different technical and scientific domains. As it happens with many NLP tasks, most computer tools are not developed to work with historical documents (cf. Quaresma and Finatto, 2020; Vieira et al., 2021; Cameron et al., 2022; Zilio et al., 2022, 2024a). As such, tools that can achieve good performance in modern data might fall short when confronted with historical writing norms. At the same time there is growing interest for the extraction of information from historical medical documents, as can be seen, among other evidences, in the recent appearance of the book *Discursos Médicos no Século XVIII* [Medical Discourses in the 18th Century] (Finatto, 2025).

In this context, this paper sets forth the task of testing a series of off-the-shelf tools to evaluate their performance in ATE using medical data written in 18th-century Portuguese. We evaluated seven tools, ranging from statistical and linguistic ATE tools to large language models (LLMs), including models trained for ATE using medical data. Far from being an attempt at a comprehensive study, this broad range of tools allowed us to start exploring the landscape of ATE for historical documents. Our test data is a single chapter of the medical handbook *Aviso a' Gente do Mar sobre a sua Saude* [Advice to Sea People about their Health], published in 1794, in which terms were manually annotated by a trained linguist.

In our exploration of ATE methods, we show results from two types of evaluation done by two linguists that are trained in the analysis of historical data: one evaluation was based on a manually annotated test data, which generated a list of target terms to be extracted, and a second evaluation was conducted based on the precision of the extraction. This second evaluation was intended to identify elements that were not considered in the original annotation, but that could help in describing important elements of the historical medical context. A third and final evaluation type arose from the com-

Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)
Vol. 2, pages 282–292

bination of these other two into a hybrid method.

The main contribution of this paper is the proposed methodologies for ATE evaluation, which include pre-annotation of a test set combined with an independent manual evaluation of the extraction. This method, while not without its faults, allowed us to greatly increase the coverage of the final extraction, and to have a larger test set.

2 Automatic Terminology Extraction

In the *Handbook of Terminology*, Heylen and De Hertog (2015, p. 203) indicate that "an expression's terminological status is often a matter of degree and open to individual variation", so ATE can help with a more objective approach to the selection of term candidates. However, continuing their argument, the authors also mention that "terms are *semantically* defined, as referring to a domain specific *concept*, and the full automatic modelling of semantics is still out of reach for computers". The handbook was written in 2015, before the development of current transformer models in NLP, and it covers ATE methods based mostly on statistics, such as collocation measures, and recurrent linguistic patterns.

Based on the results of TermEval 2020 (Terryn et al., 2020), Heylen and De Hertog (2015) seem to be right about the computer not getting a grasp of the semantics of for ATE. In the competition, a new dataset, ACTER, covering three languages (Dutch, English, and French), was released with terminological annotation, and four teams submitted their ATE tools. The winner team (for English and French) presented two BERT-based models trained on n-gram classification, which achieved f1-scores of 0.467 for English and 0.481 for French. These are very low scores to be able to reliably represent the terminology of a text. In addition, these two models were strictly language-specific.

More recently, with the further development of BERT-like transformers and the release of ALBERTINA-PT (Rodrigues et al., 2023), a new token-based classification model was developed specifically for recognizing medical terms: MediAlbertina (Nunes et al., 2024). This model was trained on named-entity recognition (NER) of medical information, and it had superior performance in comparison with other existing medical NER models, such as BioBERTpt (Schneider et al., 2020), achieving an f-score of 0.832 on the test data. This model was included in our test settings, representing the class of BERT-like, token-based classification models[1]. Because MediAlbertina was trained and fine-tuned on Portuguese data, we expected this model to be able to generalize over the slightly different historical spelling of our data.

With the current access to large language models (LLMs), Senger et al. (2025) developed a methodology to fine-tune LLMs for ATE. This methodology, Distant Supervision for Term Extraction (DiSTER), proved efficient in several datasets, but it still did not not surpass the winner of TermEval in the ACTER dataset. Because LLMs are multilingual by nature, and they are trained on huge amounts of data, they can be used for other languages, even if they were trained for extracting terms only in English. The training of the DiSTER model was also not focused on medical terms, but we expect the LLM to be able to generalize its training to the medical domain.

In this paper, we test whether the semantic definition of terms proposed by Heylen and De Hertog (2015) is "still out of reach for computers", while we acknowledge that there is still a lot of "individual variation" in what terminologists consider a term, which led us to use two complementary evaluation methods. We further complicate matters by using ATE as an umbrella term that also covers named-entity recognition in the specialized medical domain, which we will discuss in following section.

3 Of Terms and Named Entities

In the linguistic definition of terms provided by Cabré (2010, p. 357), terms are seen as "lexical units of language that activate a specialized value when used in certain pragmatic and discursive contexts. The special value results in a precise meaning recognized and stabilized within expert communities in each field". So, as much as terminology is a crucial part of our data, by itself, it would not provide us with sufficient information about the socio-historical context in which historical medical documents were produced.

This means that the information that we would like to extract from the historical sources goes beyond that of the specialized lexical units, and crosses into the territory of named entities and less-specialized lexical units. This moves our research

[1]For reference, we also tested BioBERTpt in our dataset, but we observed that the results were not satisfactory at all, possibly due to the historical spelling present in the data.

in the direction of a hybrid approach to terminology, and into the realm of a textual historical terminology, where not only, for instance, medications, diseases, and treatments are important, but also demographical and geographical data, as well as references to people that were either working on the field or being treated. Having said that, we use the word "term" throughout this paper to refer to our target units, even if we acknowledge the hybridity of our scope.

Because of this different approach to working with terminology in historical data, we could expect that some models, especially those trained on modern medical or specialized data, would fail in extracting demographical or geographical information. This is something that we take into account when prompting LLMs, as we provide them with some examples that lie outside mainstream terminology, and we observe whether they are able to adapt to this new type of information.

4 Corpus

The corpus selected for this study was originally collected in the scope of the project "Corpus Histórico da Linguagem da Medicina em Português (Séculos XVIII-XIX): Terminologia Diacrônica e Humanidades Digitais" [Historical corpus of medical language in Portuguese: Diachronic terminology and digital humanities][2] and it comprises seven chapters of the book *Aviso a' Gente do Mar sobre a sua Saude*. Its content was described in more details in the work of Zilio et al. (2024b).

The writing style in the book does not follow very strict rules for the use of terminology. One of the reasons for that could be that, by the time of writing of the document, many specialized texts were still being written in Latin, so the use of national languages for disseminating scientific knowledge and the scientific genre were still being shaped. This is also reflected in the way that terminology is used in the corpus, where a more stable, specialized vernacular lexicon was still under development for several domains. As such, instead of having precise terms and definitions, the text is written with greater fluidity, and less care for strict textual patterns. For instance, the same term referring to a "greenish stomach content" is described as "materias biliosas , e verdoengas", "materias tirante[s] a verde" e "materias biliosas , e tendentes a verde", as can be seen in these contexts, which were extracted from the test data (the highligths are ours):

	Train data	Test data	Total
Tokens	18482	2774	21256
Types	2850	923	3180

Table 1: Dataset size in types and tokens – observed with AntConc (Anthony, 2004).

> "[...] os doentes tem nauseas , vomitaõ mesmo algumas vezes espontaneamente **materias biliosas , e verdoengas** ; sua lingua se faz negra, e aspera."
> "[...] quando ha nauseas e vomitos de **materias tirante a verde**, he precizo fzer sangrias de dez para doze onças [...]"
> "[...] sobrevem-lhe desejos de vomitar ; e algumas vezes mesmo vomítos de **materias biliosas , e tendentes a verde**; todos estes symptomas chegam ao seu mais alto gráo em menos de vinte e quatro horas [...]"

Here it is important to also mention that we are dealing with a translated text. The original text was written in French by G. Mauran in 1786[3]. In 1794 it was translated and adapted by the High Surgeon of the Royal Portuguese Armada, Bernardo José de Carvalho, who took upon himself the responsibility of converting Mauran's text into a useful medical handbook for Portuguese sailors and ship surgeons. So, for instance, the three variant terminological expressions mentioned above were not variants in the original French, where Mauran consistently used "porracées" for "greenish", and even repeated the term "matières bilieuses & porracées" twice, where the Portuguese translator varied the terminology.

In this paper, we focus only on the Portuguese data, and quantitative details of the sample can be observed in Table 1. The corpus is purposefully split into train and test data, because some of the models needed input from similar information that should not come from the test data (so as to not contaminate the experiments). We thus used the train data to automatically generate linguistic patterns for the extraction with TBXTools and to extract examples of terms that were added to some of the prompts we used with LLMs.

[2]For more information about the transcription and files for download: https://sites.google.com/view/projeto38597/aviso-a-gente-do-mar-1794.

[3]The original title of the handbook was *Avis aux gens de mer, sur leur santé*.

4.1 Data Annotation

The whole corpus was annotated with terminological information using Label Studio (Tkachenko et al., 2020-2025), a Python package that provides a local Web-based annotation interface. The annotation was carried out by one linguist, who is specialized on the subject matter, but without following any annotation guidelines, except for a list of categories that were used to classify the data.

After the annotation was done, the list of terms was revised for annotation errors by the same linguist. This corpus was also annotated with morphosyntactical tags using spaCy's (Honnibal et al., 2020) *pt_core_news_lg* model.

The categories that were used as reference were the following: diseases, diagnostics, symptoms, treatments, medications, ingredients, body parts, actors, information about the population, and general medical terms. One issue arising from this list and the posterior manual analysis based on precision is that this list does not include places. So, while the manual annotation does include actors, places were left out of the list. This was one of the main sources of contributions from the system extractions. Almost all systems extracted names of places, even if not explicitly prompted to do so, because names of places can be considered as part of the information about the population. Later, in the precision-only evaluation, the extracted names of places were validated, as it can also be argued that places are an important source of information from historical data, as shown in the work of Opitz et al. (2026).

5 Methodology

Having described the corpus and the corpus annotation process, this methodology section focus on the tools that were used for (semi-)automatically extracting terms. We also dedicate some space for the second, precision-only manual evaluation process.

5.1 Off-the-shelf tools

We tested a total of seven off-the-shelf tools, some specifically developed for ATE, such as the TBXTools, DiSTER and the MediAlbertina models, and others that were developed with different purposes in mind, but that have features that allow for their use in ATE, such as the Sketch Engine, which provides, among several other features, keyword and term extraction tools, and the MedGemma, Gemma and EuroLLM models, which were developed to complete a text input provided by a user, but that can be prompted to extract terminology from data. Some of the tools mentioned in this subsection were already partially introduced in Section 2, so we will not expand on them as much as on the others.

5.1.1 Pattern-based extraction models

TBXTools (Oliver and Vàzquez, 2015). This tool presents two options for ATE: a linguistic extraction, and a statistical extraction based on n-gram sizes and, potentially, association measures. Because of the small size of the corpus, the statistical extraction feature was not used. The tool provides the means to automatically extract linguistic patterns based on existing annotated data, so we used our train data to extract morphsyntactical patterns. These extracted patterns were then manually revised and cleaned, so this was a semi-automatic extraction. The cleaning was necessary, because the morphosyntactical annotation done by spaCy is not very precise on historical data, and several patterns were spurious. Even with the cleaned patterns, there were still some basic issues with the ATE, as the corpus contained some errors in the recognition of stopwords such as "he" [is], "naõ" [no / not], which were sometimes mistakenly annotated, for instance, as nouns. This happens because of the historical spelling of these words. Having this in mind, after we used the cleaned patterns to extract term candidates from the test data, we used a post-processing python script to ensure that no term candidate would begin or end with stopwords. More details about this process are presented further down in this subsection. The semi-automatic extraction with TBXTools resulted in 331 terms that ranged from bigrams to heptagrams.

Sketch Engine (SkE) (Kilgarriff et al., 2014). Sketch Engine provides two types of extraction: keywords and terms. As such, contrary to TBXTools, SkE can also extract unigram terms. Although keywords and terms have different theoretical implications (the former being salient tokens in a text, and the latter being the carriers of specialized concepts in a given domain), there is usually a good amount of overlap between the two in specialized domains, so we treated the results of the keyword extraction as ATE. Similarly to the TBXTools, for terms longer than unigrams, SkE uses linguistic patterns for ATE, and so it also relies on its own morphosyntactical annotation to extract word combinations that satisfy previously

conceived terminological patterns. As such, for the same reasons explained for TBXTools, we applied a cleaning script at the end, to ensure that no extracted term candidate was a stopword, or had a stopword at its beginning or end. Because SkE also included unigram extraction, it provided us with more term candidates: 873 in total (724 unigrams).

Stopwords. As mentioned for both SkE and TBXTools, we used a custom list of stopwords to remove terms that either were a stopword (in the case of SkE), or started or ended with a stopword. To create this custom list, we extracted the first 150 most common tokens from the train corpus, manually removed a few non-stopwords, and manually added a few singular or plural forms of grammatical words that were already present, such as "elle" [he], "ás" [to the], etc. The resulting list contained 124 words, which were merged with a larger list of contemporary Portuguese stopwords[4]. The final list contained 592 unique items. In addition, due to slight differences in historical orthography, we also replaced all instances of "ão" in the stopword list with "aõ", to increase its coverage.

5.1.2 Large language models (LLMs)

We tested two types of large language models: specialized models, which were fine-tuned for ATE or for working with medical data, and generic models, which have no extra fine-tuning.

DiSTER (Senger et al., 2025). The *DiSTER-Llama-3-8B-Instruct*[5] model is a fine-tuned version of Llama-3-8B model. This is a generative LLM that was specifically fine-tuned for ATE using a combination of datasets from several domains, including biomedicine. All datasets used in the model's fine-tuning were in English, but Llama's multilingual nature was able to provide results for Portuguese as well. We used the default parameters as suggested on the model's Huggingface page.

Gemma 3 (Gemma Team, 2025). The model *Gemma-3-4B-It*[6] is a generic LLM, trained to complete a prompt with the most probable tokens. Only instruction fine-tuning was applied to this model, so we relied solely on the prompting strategies to guide it to perform ATE. Regarding the default parameters presented on the model's Huggingface page, we made two changes: *temperature* was set to 0.0, and *max_new_tokens* was set to 1024.

MedGemma (Sellergren et al., 2025). *MedGemma-4B-It*[7] is a Gemma model that was fine-tuned over several medical datasets, but not necessarily for ATE. Similarly to what we did with Gemma 3, for MedGemma we also only changed two parameters: *temperature* was set to 0.0, and *max_new_tokens* was set to 1024.

EuroLLM (Martins et al., 2025). *EuroLLM-9B-Instruct*[8] is a generic LLM trained on 35 languages. Similarly to Gemma, MedGemma, and DiSTER, it is an instruction-tuned model, which help in our task of prompting it with instructions to extract terminology. Again, we did not perform many changes to the model's parameters in regard to the default settings presented on the Huggingface page, but we did set its *temperature* to 0.01.

Prompting LLMs for ATE. We used four prompting strategies with each of the four models, and thus got 16 different results for the LLM extraction. Due to the slightly different training of the models, the prompts were also slightly different in terms of structure, but the content of the actual prompts was the same. Each prompt also contained a paragraph of the test data, with an average size of 132.29 tokens. There were 24 paragraphs in the test data, so each model was prompted 24 times for each prompting strategy (*i.e.*, 96 times per model). All prompts were written in English, which might have helped in highlighting the paragraphs written in Portuguese as the target of the task.

The strategies used were the following, increasing in information at each step:

- **Zero shot**: this strategy is the one that provides the least information to the system. Apart from mentioning the medical domain and the ATE task, no other information was provided, fully relying on the model's capacity to generate lists of terms.

- **Categories**: in addition to providing information about the domain, in this prompt, a few categories of interest were presented to the model: diseases, diagnostics, symptoms, treatments, medications, ingredients, body parts, actors, information about the population, and general medical terms.

[4]The larger, contemporary Portuguese stopword list was downloaded from https://github.com/stopwords-iso/stopwords-iso.

[5]https://huggingface.co/ElenaSenger/DiSTER-Llama-3-8B-Instruct.

[6]https://huggingface.co/google/gemma-3-4b-it.

[7]https://huggingface.co/google/medgemma-4b-it.

[8]https://huggingface.co/utter-project/EuroLLM-9B-Instruct.

- **One shot**: in this prompt, in addition to the categories, we also provided the system with a single example-term, which was not present in the test data.
- **Few shots**: here we gave the model one example-term per category. None of the example-terms were present in the test data.

The prompts used for each system are presented in more detail in Appendix A.

MediAlbertina (Nunes et al., 2024). MediAlbertina was created by fine-tuning Albertina PT (Rodrigues et al., 2023) for medical named-entity recognition. This is the only model that was trained exclusively on Portuguese data, and later fine-tuned on medical data written in Portuguese. Contrary to the previous generative models, whose task is to complete a prompt, MediAlbertina's task is to classify tokens of a text in a BIO fashion, so it evaluates each token in a text as a potential candidate, and outputs a label, either as a **b**eginning or **i**nternal part of a named entity, or as being **o**utside of named entities (i.e. as not being part of a named entity). This classification is then retrieved as a list of terms (or, to be more precise, as a list of medical named entities).

5.2 Evaluation

Three types of evaluation were carried out: a manual precision-only evaluation, an automatic f-measure evaluation, and an automatic hybrid evaluation.

In the precision-only evaluation, the list of extracted term candidates was analyzed by a linguist, and the candidates were classified as terms, non-terms, or partial terms. The classification was supported by contexts of occurrence (*i.e.*, the paragraphs from where candidates were extracted), and the occurrence as a term in at least one context would suffice for the classification of the candidate as a valid term. In total, 3,208 candidates were evaluated this way. The initial evaluation was semi-automatically revised by a second linguist, focusing on cases of inter-model annotation disagreements.

The second evaluation method, an f-measure evaluation, was carried out automatically, based on the annotated test data, as detailed in Subsection 4.1. The test data contained 193 terms. When contrasted with the precision-only annotation, another 252 unique terms were added to this list, which was then used in our final evaluation method: a hybrid evaluation, combining the annotated terms with the precision-only terms that were evaluated as valid. This goes to show that a single annotator, who is specialized in the topic, was able to cover around 43.37% of the information. This percentage rises to 61.12% if we consider that 79 of the new terms were part of the 193 annotated terms (that is, out of the 252 added terms, 79 were actually part of a longer term already present in the annotated data). This happened because, during the annotation process, the longest term would be selected, and any internal, shorter terms would not be individually annotated. We also have to consider that there are terms in the data that were potentially left out by the human annotator and by all the systems as well.

This process allowed us to observe not only the quality of the models used for ATE, but also the coverage that can be expected from human annotation, and how it can be improved by adding automatic tools in the process.

6 Results and Discussion

Table 2 presents the results divided by type of model: pattern-based extraction models (PB model), generative artificial intelligence models (GenAI models) and the single token-based classification model (TBC model). For each of the GenAI models, we also included subscripted information about the type of prompt used for obtaining the indicated results: *zero* stands for zero-shot prompting, *ner* represents the category-based approach, *one* refers to the one-shot approach, and *few* indicates the few-shot approach.

The table also contains information about how many unique term candidates each model extracted, and about how many new validated terms each model contributed to the hybrid evaluation. In the precision-only evaluation, the terms that were considered as partially correct were considered as correct in the "lenient precision" column. The top results for each evaluation category are highlighted in bold. There is no information about significance, because all models are deterministic or very close to deterministic[9].

The Gemma-family of models provided the best f-measure scoring, when we consider the hybrid evaluation (*i.e.*, annotation + precision-only). The

[9]The GenAI models, even at very low or zero temperature, still present minor fluctuations in their outputs, which are explained in (He and Lab, 2025). We consider that these minor fluctuations would not suffice to warrant several prompting attempts, in order to establish a mean and standard deviation as proposed by (De Pourcq et al., 2025).

Models	Precision-only Evaluation			Annotation Evaluation			Hybrid Evaluation			
	Extracted Candidates	Strict Precision	Lenient Precision	Precision	Recall	f-measure	New Terms	Precision	Recall	f-measure
PB models:										
Sketch Engine	873	0.2726	0.4742	0.1661	**0.7513**	0.2720	102	0.3253	**0.6382**	0.4310
TBXTools	331	0.4139	0.6344	0.1722	0.2953	0.2176	83	0.4381	0.3258	0.3737
GenAI models:										
DiSTER$_{zero}$	156	**0.7244**	**0.7821**	0.5000	0.4041	**0.4470**	38	**0.7564**	0.2652	0.3927
DiSTER$_{ner}$	99	**0.7172**	0.7677	**0.5556**	0.2850	0.3767	18	**0.7475**	0.1663	0.2721
DiSTER$_{one}$	97	**0.7629**	**0.7835**	**0.5876**	0.2953	0.3931	18	**0.7732**	0.1685	0.2768
DiSTER$_{few}$	128	0.6094	0.7109	0.4688	0.3109	0.3738	22	0.6719	0.1933	0.3002
EuroLLM$_{zero}$	357	0.3950	0.5854	0.2241	0.4145	0.2909	67	0.4370	0.3506	0.3890
EuroLLM$_{ner}$	365	0.4164	0.6082	0.2055	0.3886	0.2688	82	0.4548	0.373	0.4099
EuroLLM$_{one}$	320	0.4531	0.6344	0.2500	0.4145	0.3119	72	0.5031	0.3618	0.4209
EuroLLM$_{few}$	354	0.3927	0.5791	0.2062	0.3782	0.2669	74	0.4407	0.3506	0.3905
Gemma$_{zero}$	348	0.4511	0.6178	0.2759	0.4974	0.3549	66	0.4856	0.3798	0.4262
Gemma$_{ner}$	287	0.6132	**0.7909**	0.3659	0.5440	0.4375	76	0.6516	0.4202	**0.5109**
Gemma$_{one}$	285	0.6070	0.7719	0.3895	0.5751	**0.4644**	69	0.6526	0.418	**0.5096**
Gemma$_{few}$	260	0.6462	**0.8000**	0.3962	0.5337	**0.4547**	70	0.6808	0.3978	**0.5021**
MedGemma$_{zero}$	618	0.2670	0.4337	0.1553	0.4974	0.2367	80	0.3026	0.4202	0.3518
MedGemma$_{ner}$	365	0.5370	0.6959	0.3288	**0.6218**	**0.4301**	84	0.5863	**0.4809**	**0.5284**
MedGemma$_{one}$	344	0.5581	0.7413	0.3372	**0.6010**	**0.4320**	83	0.6017	**0.4652**	**0.5247**
MedGemma$_{few}$	389	0.4730	0.6247	0.2725	0.5492	0.3643	84	0.5090	0.4449	0.4748
TBC model:										
MediAlbertina	18	**0.7778**	**0.9444**	**0.5000**	0.0466	0.0853	5	**0.7778**	0.0315	0.0605

Table 2: Results of the manual precision-only evaluation, of the automatic f-measure evaluation based on the manually annotated test set, and of the automatic hybrid evaluation.

non-adapted version had slightly lower scores, but, given some basic information, such as the categories of interest, and possibly a single example, these models were able to achieve a good balance of number of extracted candidates, precision and recall. If the aim of the task is, however, to extract fewer candidates that are more precise, then the DiSTER model toped the table, achieving up to 77.32% precision, but with an extraction of only 97 terms. Usually, however, in tasks like ATE, the focus is on the extraction of as many terms as possible, and, in terms of recall, no system was better than SkE. It did extract a lot of candidates, which warranted the lowest score in precision in the trade-off, but, with a score of 63.82%, it toped the list in recall.

Considering that no system by itself achieved a high score, we tested combinations of two and three models, focusing on improving precision, recall and f-measure with two models, and then purely recall and f-measure with three models. These combinations were tested directly on the hybrid test set, that is, the one that emerged from the combination of the two manual evaluations. Tables 3 and 4 show results for combinations of two and three models, respectively. For the lack of space, not all possible permutations are shown in these tables, which focus on the most relevant results. In Table 4 we did not highlight any results for precision, because the best results in the combination of three models was 74.82% over 139 unique terms, when combining DISTER$_{ner}$ + DISTER$_{one}$ + MediAlbertina, but this result was already inferior to DISTER$_{one}$ by itself (see Table 2), and f-measure of the combo was much lower than the other combinations, at 35.62%.

The combinations of different models was a very promising path to explore. It is expected that, by joining different models, some will be complementary and achieve better results together, especially in terms of recall and f-measure. Here we could see that the combination of the two pattern-based extraction models, SkE and TBXTools toped the table in recall, even if TBXTools did not excel in any front, it was the only model that complemented SkE's extraction in such a way as to increase recall by almost 20 percentage points. Precision was not very high, as both models together extracted 1121 unique terms, but the f-measure was much better than many models by themselves. And when these two models were joined by MedGemma or by EuroLLM, the recall jumped above 90 percentage points, while still keeping an f-measure at around 48 points.

Combination	Unique Terms	Precision+	Recall+	f-measure+
DiSTER$_{ner}$ + Gemma$_{few}$	298	0.6443	0.4315	0.5168
DiSTER$_{one}$ + Gemma$_{few}$	298	0.6510	0.4360	0.5222
DiSTER$_{zero}$ + Gemma$_{few}$	317	0.6562	0.4674	0.5459
DiSTER$_{zero}$ + MedGemma$_{ner}$	411	0.5888	0.5438	**0.5654**
EuroLLM$_{ner}$ + TBXTools	685	0.4409	0.6787	0.5345
Gemma$_{few}$ + MediAlbertina	270	**0.6815**	0.4135	0.5147
Gemma$_{ner}$ + TBXTools	579	0.5095	0.6629	**0.5762**
Gemma$_{one}$ + MediAlbertina	291	0.6564	0.4292	0.5190
MedGemma$_{few}$ + TBXTools	675	0.4519	0.6854	0.5446
MedGemma$_{ner}$ + TBXTools	641	0.4805	0.6921	**0.5672**
MedGemma$_{one}$ + TBXTools	617	0.4830	0.6697	**0.5612**
SkE + TBXTools	1121	0.3318	**0.8360**	0.4751

Table 3: Results for model combinations using the hybrid evaluation test set.

Combination	Unique Terms	Precision+	Recall+	f-measure+
DiSTER$_{zero}$ + Gemma$_{ner}$ + TBXTools	627	0.5088	0.7169	**0.5951**
DiSTER$_{zero}$ + Gemma$_{one}$ + TBXTools	623	0.5072	0.7101	**0.5918**
EuroLLM$_{ner}$ + SkE + TBXTools	1253	0.3256	**0.9169**	0.4806
EuroLLM$_{one}$ + SkE + TBXTools	1230	0.3285	**0.9079**	0.4824
MedGemma$_{ner}$ + SkE + TBXTools	1213	0.3331	**0.9079**	0.4873
MedGemma$_{few}$ + SkE + TBXTools	1242	0.3285	**0.9169**	0.4837
MedGemma$_{one}$ + SkE + TBXTools	1215	0.3342	**0.9124**	0.4892

Table 4: Results from the combination of three models using the hybrid evaluation test set.

6.1 Of Models and Hallucinations

It is known that GenAI models can produce spurious results, usually referred to as hallucinations. They can generate outputs that are not real, or that do not correspond to the given task. What we saw in our data was that, apart from TBXTools and MediAlbertina, which would only produce hallucinations in the form of false positives, all the models had their own ways of producing hallucinations, not only the the LLMs.

The Gemma-family models sometimes "corrected" words present in the document. For instance, "vomitos" would sometimes be modernized to "vômitos" [vomit] in the extraction, "emulssaõ" was modified to "emulssa", which is not a word in Portuguese, and "respiraõ" was modified to "respira", a different form of the verb "respirar" [to breathe]. These modifications were accepted in the lenient precision-only evaluation (as partial matches), but not in the f-measure calculations.

EuroLLM also produced alterations in the spelling of extracted data, similar to Gemma. In addition, it would sometimes get into a loop, where it would repeat a word up to the maximum number of generated tokens. That's why it featured, for instance, the pronoun "tudo" [everything] 223 times, and "primeira" [first$_{feminin}$] 230 times in its outputs.

The DiSTER model frequently extracted information directly from the prompt, instead of extracting term candidates only from the target paragraph, especially in cases where the target paragraphs did not have any terms to be extracted. As such, the output would frequently contain the categories of interest (preserved in English) or the example-terms indicated in the prompt (even if they were not present in the test data). The zero-shot model was the only one not to produce such outputs.

SkE does not generate new output by itself, but, because it relies on lemmatized data to extract terms, and because its lemmatizer is not trained on historical Portuguese data, some candidates were extracted with bad grammatical agreement, such as "alimento mais succosos" [juicier$_{plural}$ food] and "remedios administrado" [administered$_{singular}$ medications]. As it happened with Gemma, when

these could be considered terms, they were considered correct in the lenient precision-only evaluation, but not when calculating the f-measure.

Even though MedGemma and EuroLLM had some hallucinations in the same way as the other LLMs did, they also provided some interesting results that we were not expecting to see. They were able to leverage their generating powers to combine words that were separated in the text, but that belonged together as a term. For instance, in the context "o ventre esta tenso , e dorido" [the abdomen is tense, and hurting], EuroLLM was able to extract "ventre dorido", and, in the context "quando a sede he urgente" [when the thirst is pressing], MedGemma joined together "sede urgente". These cases were rare, with five occurrences for EuroLLM and three for MedGemma, but they were considered as correct extractions and were added to the hybrid f-measure evaluation. DiSTER also showed potential for doing this, but it happened only twice, in the same context, as it extracted "vinho de Alicante" and "vinho de Chipre" from the context "algumas colheres do vinho de Malga , de Chipre , de Alicante" [some spoons of Malaga, Cyprus, or Alicante wine].

7 Final Remarks

In general, the models' ATE performance was not bad. We cannot draw a direct parallel with TermEval data, as we are working with a completely different dataset and evaluation method, but the f-score of 0.5951 for the combination of DiSTER$_{zero}$ + Gemma$_{ner}$ + TBXTools can be taken as a promising result. However, it was a surprise to see that, even with all the developments in neural and GenAI models, the pattern-based models still performed better in recall, which is arguably the most important metric for a terminologist, and arguably also the most important metric for us in the analysis of historical information. TBXTools and SkE, if not very performant by themselves, provided a great complement to one another and to other GenAI models, as they were the most recurrent models in the best combinations.

The use of two evaluation methods, which then generated a third evaluation, seemed to be a good approach for this experiment. By combining a token-based annotation with a precision-only evaluation, we were able to highlight how much models can contribute to a human annotation, and to show that neither a single human nor a single model can achieve high f-measure in ATE, even if the human is more precise in their annotations.

The outputs of GenAI models also presented some cause for concern, as EuroLLM and both Gemma-based models produced alterations in the spelling extracted candidates. Alterations and modernizations of spelling can have a significant impact in the description of historical texts and in the compilation of terminologies that were still being consolidated at the time.

Overall, the results of the experiments reported in this paper give us more confidence moving forward to the analysis of the remaining chapters in the dataset. By focusing on automatically extracted data, combined with a more detailed human analysis of the contexts, we can extract valuable information from the historical medical data that can be used to describe past medical practices and to further advance the field of Digital Humanities.

Limitations

One of the main limitations of this paper was its scope. Due to the amount of data that needed to be annotated and evaluated, we could not work with a larger dataset, and had to settle for a single chapter of the medical handbook as test sample.

A second limitation was the inexistence of annotation guidelines or of more annotators. We could not evaluate, for instance, how human annotators would agree on the annotation of terminological information that goes beyond the more strict boundaries of terms.

Perhaps more a trade-off than an actual limitation was the slight difference, in comparison to the literature, in the way we evaluated the tools. We did not perform a token-based evaluation, but rather focused on a type-based evaluation, where lists of terms were compared, instead of lists of frequencies or precise token positions. This approach was different from what is usually found in the literature. For instance, in Terryn et al. (2020), each occurrence of a type was taken into account for calculating the f-score. In our evaluation, even a single automatic extraction of a term that could occur ten times in the test set would result in an f-score of 1 for that term, but it also meant that terms that are very frequent would be treated in the same way as rarer terms. In this way, a system that would detect multiple occurrences of a frequent term, while letting slide a rarer, single-occurrence term, would be penalized by 50%.

Acknowledgments

This research was supported in Belgium by the Wallonia-Brussels Federation's Special Research Fund (ILC FSR24) and in Brazil by the National Council for Scientific and Technological Development (CNPq), grant PQ 307088/2023-5, PIBIC-CNPq-UFRGS, FAPERGS - Edital 06/2025, and TILD-IAR-CNPq (grant 408490/2024-1).

References

Laurence Anthony. 2004. Antconc: A learner and classroom friendly, multi-platform corpus analysis toolkit. pages 7–13.

M Teresa Cabré. 2010. Terminology and translation. *Handbook of translation studies*, 1:356–365.

Helena Cameron, Fernanda Olival, Renata Vieira, and Joaquim Santos. 2022. Named entity annotation of an 18th-century transcribed corpus: problems and challenges. In *Proceedings of the Second Workshop on Digital Humanities and Natural Language Processing (2nd DHandNLP 2022) co-located with International Conference on the Computational Processing of Portuguese (PROPOR 2022), Fortaleza, Brazil, 21st March, 2022*, pages 18–25. CEUR.

Lena De Pourcq, Marie Gregoire, and Leonardo Zilio. 2025. Exploring the power of generative artificial intelligence for automatic term extraction from small samples. In *Electronic lexicography in the 21st century (eLex 2025) Intelligent Lexicography. Proceedings of the eLex 2025 conference*, pages 116–138. Lexical Computing CZ s.r.o.

Maria José Bocorny [Org.] Finatto. 2025. *Discursos Médicos no Século XVIII: genealogia de saberes e conhecimentos através da linguagem*. Editora da ABRALIN.

Gemma Team. 2025. Gemma 3.

Horace He and Thinking Machines Lab. 2025. Defeating nondeterminism in llm inference. *Thinking Machines Lab: Connectionism*. Https://thinkingmachines.ai/blog/defeating-nondeterminism-in-llm-inference/.

Kris Heylen and Dirk De Hertog. 2015. Automatic term extraction. *Handbook of terminology*, 1(01).

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spacy: Industrial-strength natural language.

Adam Kilgarriff, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlỳ, and Vít Suchomel. 2014. The sketch engine. *Lexicography*, 1(1):7–36.

Pedro Henrique Martins, João Alves, Patrick Fernandes, Nuno M Guerreiro, Ricardo Rei, Amin Farajian, Mateusz Klimaszewski, Duarte M Alves, José Pombal, Nicolas Boizard, and 1 others. 2025. Eurollm-9b: Technical report. *arXiv preprint arXiv:2506.04079*.

Miguel Nunes, João Boné, João C Ferreira, Pedro Chaves, and Luis B Elvas. 2024. Medialbertina: an european portuguese medical language model. *Computers in Biology and Medicine*, 182:109233.

Antoni Oliver and Mercè Vàzquez. 2015. Tbxtools: A free, fast and flexible tool for automatic terminology extraction. In *Proceedings of the international conference recent advances in natural language processing*, pages 473–479.

Juri Opitz, Corina Raclé, Emanuela Boros, Andrianos Michail, Matteo Romanello, Maud Ehrmann, and Simon Clematide. 2026. Clef hipe-2026: Evaluating accurate and efficient person-place relation extraction from multilingual historical texts. *arXiv preprint arXiv:2602.17663*.

Paulo Quaresma and Maria José Bocorny Finatto. 2020. Information extraction from historical texts: a case study. In *Proceedings of the Workshop on Digital Humanities and Natural Language Processing, co-located with International Conference on the Computational Processing of Portuguese, DHandNLP@PROPOR, Evora, Portugal, March 2, 2020.*, pages 49–56. CEUR.

João Rodrigues, Luís Gomes, João Silva, António Branco, Rodrigo Santos, Henrique Lopes Cardoso, and Tomás Osório. 2023. Advancing neural encoding of portuguese with transformer albertina pt. In *EPIA Conference on Artificial Intelligence*, pages 441–453. Springer.

Elisa Terumi Rubel Schneider, João Vitor Andrioli de Souza, Julien Knafou, Lucas Emanuel Silva e Oliveira, Jenny Copara, Yohan Bonescki Gumiel, Lucas Ferro Antunes de Oliveira, Emerson Cabrera Paraiso, Douglas Teodoro, and Cláudia Maria Cabral Moro Barra. 2020. BioBERTpt - a Portuguese neural language model for clinical named entity recognition. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, pages 65–72, Online. Association for Computational Linguistics.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, and 1 others. 2025. Medgemma technical report. *arXiv preprint arXiv:2507.05201*.

Elena Senger, Yuri Campbell, Rob Van Der Goot, and Barbara Plank. 2025. Crossing domains without labels: Distant supervision for term extraction. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1366–1378, Suzhou (China). Association for Computational Linguistics.

Ayla Rigouts Terryn, Veronique Hoste, Patrick Drouin, and Els Lefever. 2020. Termeval 2020: Shared task on automatic term extraction using the annotated corpora for term extraction research (acter) dataset. In *Proceedings of the 6th International Workshop on Computational Terminology*, pages 85–94.

Maxim Tkachenko, Mikhail Malyuk, Andrey Holmanyuk, and Nikolai Liubimov. 2020-2025. Label Studio: Data labeling software. Open source software available from https://github.com/HumanSignal/label-studio.

Renata Vieira, Fernanda Olival, Helena Cameron, Joaquim Santos, Ofélia Sequeira, and Ivo Santos. 2021. Enriching the 1758 portuguese parish memories (Alentejo) with named entities. *Journal of Open Humanities Data*, 7:20.

Leonardo Zilio, Maria Finatto, and Renata Vieira. 2022. Named entity recognition applied to Portuguese texts from the XVIII century. In *Proceedings of the Second Workshop on Digital Humanities and Natural Language Processing (2nd DHandNLP 2022) co-located with International Conference on the Computational Processing of Portuguese (PROPOR 2022), Fortaleza, Brazil, 21st March, 2022*, pages 1–10. CEUR.

Leonardo Zilio, Rafaela R Lazzari, and Maria José B Finatto. 2024a. Can rules still beat neural networks? The case of automatic normalisation for 18th-century Portuguese texts. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese-Vol. 2*, pages 83–92.

Leonardo Zilio, Rafaela Radünz Lazzari, and Maria Jose Bocorny Finatto. 2024b. NLP for historical Portuguese: Analysing 18th-century medical texts. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese-Vol. 1*, pages 76–85.

A Query Structures and Prompts

There were, in total, three different query structures and four prompting strategies. So, for instance, the query structure used for the DiSTER model was the following:

```
'{"id": "test_0", "conversations": [\n'
'    {"from": "human", "value": "Text: ' + processed_line + '"},\n'
'    {"from": "gpt", "value": "I\'ve read this text."},\n'
'    {"from": "human", "value": "' + prompt + '"},\n'
'    {"from": "gpt", "value": ""}\n'
']}\n'
```

Where "processed_line" was a variable containing the current paragraph of the test set, and "prompt" was a variable containing the following string, for the zero-shot strategy:

> Extract a single list of medical terms from a short text. The terms can range from unigrams to n-grams. The output should only contain the list of terms, in a format that can be directly read as a Python list.

This query structure followed the one available on the model card on Huggingface.

For EuroLLM and the Gemma-family models, the query was very similar, with just a slight change. Here is the query structure for EuroLLM as an example:

```
{
    "role": "system",
    "content": "You are an expert terminologist that works with \
        18th-century medical documents written in Portuguese.",
},
{
    "role": "user",
    "content": current_prompt
}
```

Again, the "current_prompt" here contained the full prompt, such as this, for zero shot:

> Extract a single list of medical terms from a short text. The terms can range from unigrams to n-grams. The output should only contain the list of terms, in a format that can be directly read as a Python list. Here is the text: '##########'

Where "##########" is just a placeholder that would be replaced with the actual paragraph from the test set.

As these examples show, the queries for DiSTER were slightly different due to the different query structure, but the content of the instruction prompt was the same.

For approaches with information about categories, the following string was added after "n-grams":

> , and might include diseases, diagnostics, symptoms, treatments, medications, ingredients, body parts, actors, information about the population, and general medical terms.

This was further complemented with:

> Example of term (not included in the sample): peripneumonia.

for the one-shot strategy, and

> Examples of terms (not included in the sample): peripneumonia, pulsaçaõ, amollecido, panada nutriente, mel, intestinos, Professores, robusto, tratamento.

for the few-shot strategy.

Marcação semântica de entidades nomeadas em *Os Lusíadas*

Adriane Maria de Oliveira Queiroz
Universidade Federal da
Grande Dourados (UFGD)
Programa de Pós-Graduação
em Letras (PPGL)
Dourados, MS, Brasil
adrianeoliqueiroz@gmail.com

Bruno Oliveira Maroneze
Universidade Federal da
Grande Dourados (UFGD)
Faculdade de Comunicação,
Artes e Letras (FALE)
Dourados, MS, Brasil
brunomaroneze@ufgd.edu.br

Abstract

Este artigo apresenta a modelagem semântica de entidades nomeadas em *Os Lusíadas*, de Luís de Camões, com base no padrão TEI P5. Propõe-se um fluxo híbrido de anotação que combina NER (spaCy), dicionário de autoridade (gazetteer) e pós-edição filológica manual. São tipificados antropônimos, mitônimos e topônimos por meio dos elementos `<persName>` (nome de pessoa), `<placeName>` (nome de lugar) e `<rs>` (referencing string, para cadeias de referências), com especial atenção à marcação de epítetos. O estudo evidencia os limites de modelos treinados em corpora jornalísticos diante da sintaxe épica e da ortografia da edição de 1572, demonstrando a necessidade de uma abordagem híbrida. Conclui-se que o XML/TEI atua como ferramenta de modelagem do conhecimento literário.

1 Introdução

A preservação e o estudo de monumentos literários como *Os Lusíadas*, de Luís de Camões, têm passado por uma mudança de paradigma: a transição da digitalização meramente visual para a modelagem profunda de dados textuais. No entanto, a aplicação de técnicas de Processamento de Linguagem Natural (NLP) e Reconhecimento de Entidades Nomeadas (NER) em textos do século XVI impõe obstáculos severos. A densidade onomástica da épica camoniana, caracterizada por uma complexa rede de referências históricas e mitológicas, aliada à instabilidade ortográfica da edição de 1572 e à sintaxe latinizante, cria um cenário de *domain shift* (desvio de domínio), situação na qual modelos contemporâneos de NLP treinados em corpora jornalísticos apresentam desempenho subótimo.

Este trabalho, integrante do projeto "Lusíadas Digital", descreve a implementação de um fluxo de trabalho (*workflow*) híbrido e "human-in-the-loop" para a anotação semântica de antropônimos, mitônimos e topônimos em *Os Lusíadas*. A pesquisa não visa apenas a extração automática de dados, mas a construção de uma infraestrutura de conhecimento baseada no padrão TEI P5 (Text Encoding Initiative) (TEI Consortium, 2023). A inovação reside na integração harmônica entre a escala computacional e o rigor da tradição filológica luso-brasileira, utilizando o XML/TEI como uma ponte dialógica entre a crítica textual clássica e os métodos quantitativos das Humanidades Digitais. O tratamento lexical e onomástico do corpus dialoga com a tradição crítica camoniana e filológica, particularmente com os comentários e estudos de autores como Epiphanio da Silva Dias (Camões, 1916) e Aguiar e Silva (2010). O objetivo final é demonstrar como a colaboração entre algoritmos de inteligência artificial e a curadoria especializada permite converter o poema em um grafo de conhecimento interoperável e dinâmico.

2 Fundamentação Teórica e Modelo Semântico

A presente pesquisa situa-se na interseção entre Filologia Digital, Humanidades Digitais e Processamento de Linguagem Natural, partindo da premissa de que a modelagem textual não é apenas um procedimento técnico, mas uma operação interpretativa. A codificação em TEI P5 não representa somente a estrutura formal do texto, mas também explicita hipóteses críticas sobre identidade, referência e função narrativa.

A definição do modelo semântico em um projeto de Filologia Digital exige o que a ciência da computação denomina "compromisso ontológico": a decisão de quais categorias da realidade serão

Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)
Vol. 2, pages 293–297

representadas e como serão hierarquizadas. No caso d'*Os Lusíadas*, a escolha recaiu sobre o uso dos elementos <persName>, <placeName> e <rs>, todos pertencentes ao módulo namesdates da TEI P5. Esta escolha, contudo, não é a única possível e convida a uma reflexão sobre a granularidade e a finalidade da marcação.

Uma alternativa simplificada seria o uso da tag genérica <name>. Embora o elemento <name> reduza a complexidade da anotação automática, ele falha ao não distinguir semanticamente a natureza da entidade. Para um poema que transita entre o registro histórico e o maravilhoso pagão, a distinção entre <persName> e <placeName> é vital para futuras extrações de dados.

Dentro de <persName>, optou-se pela tipificação via atributo @type ("hist" para personagens históricos e "myth" para divindades). Discute-se na comunidade TEI se divindades deveriam possuir uma tag própria (como uma hipotética <mythName>), mas a prática consensual reforça que, funcionalmente, deuses atuam como actantes (Greimas, 1973) e pessoas no discurso épico. Em Semântica estrutural, Greimas propõe que a narrativa pode ser descrita a partir de funções estruturais abstratas (sujeito, objeto, destinador, destinatário, adjuvante e oponente) que independem da materialidade lexical do personagem. Essa distinção entre personagem empírico e função narrativa é fundamental para a modelagem digital, pois permite compreender que diferentes expressões linguísticas podem remeter à mesma instância actancial. Assim, a distinção por atributo em vez de elemento mantém a compatibilidade com ferramentas de análise de redes sociais (Social Network Analysis) (Moretti, 2011), que buscam interações entre "pessoas", independentemente de sua natureza metafísica.

O ponto mais sensível do modelo é a marcação de referências indiretas. Camões utiliza a antonomásia como recurso estilístico central (ex: "O forte Capitão" para Vasco da Gama). Existem duas vias de marcação aqui:

1. Marcar "O forte Capitão" como um nome de pessoa.

2. Marcar como uma "cadeia de referência".

A identificação automática de entidades nomeadas (NER) reconhece apenas parte desse fenômeno, já que modelos estatísticos tendem a privilegiar formas canônicas. A crítica textual, por sua vez, evidencia que expressões como "O forte Capitão" operam como marcadores identitários plenos, ainda que não apresentem um nome próprio explícito. Nesse sentido, a adoção do elemento <rs> no padrão TEI permite registrar cadeias referenciais que extrapolam a simples nomeação. A distinção entre <persName> e <rs> torna-se metodologicamente relevante: enquanto a primeira fixa uma entidade tipificada, a segunda preserva a dimensão retórica da enunciação.

Defendemos que o uso de <rs> é a abordagem superior. Marcar um epíteto como <persName> constitui um erro filológico, pois confunde o "nome de batismo" com a caracterização literária. O elemento <rs>, aliado ao atributo @ref, permite que o pesquisador mapeie a fama da personagem, ou seja, como ela é construída e referenciada indiretamente ao longo dos cantos, sem corromper a taxonomia dos nomes próprios. Essa escolha possibilita, por exemplo, o estudo estatístico da frequência com que Camões evita o nome próprio em favor do título épico.

Para os lugares, o uso de <placeName> com o atributo @type="real" ou "myth" resolve a dicotomia geográfica predominante no poema. Todavia, em versos de alta densidade erudita, o modelo enfrenta o desafio da polissemia. O verso "O Pado o sabe, o Lampetusa o sente" (I, 46) é o exemplo mais eloquente dessa complexidade, onde a marcação semântica exige o respaldo do aparato crítico tradicional para ser precisa.

A decisão de marcar "Lampetusa" apenas como um lugar geográfico (a ilha de Lampedusa), como é marcada automaticamente, ignoraria o "gesto interpretativo" inerente à camonologia. O *Dicionário d'Os Lusíadas* de Afrânio Peixoto e Pedro A. Pinto (1924) (Peixoto and Pinto, 1924), bem como o *Dicionário e Gramática de Os Lusíadas* de Júlio Nogueira (1960) (Nogueira, 1960), são categóricos ao definir Lampetusa como uma das Helíades, as irmãs de Faetonte, que choraram a morte do irmão às margens do rio Pado. O desafio de codificação amplia-se ao considerar a nota filológica de Augusto Epiphanio da Silva Dias (Camões, 1916), em sua versão comentada. Silva Dias observa que a escolha do nome por Camões reflete uma linhagem de fontes específicas: enquanto Ovídio nomeia Phaethusa e Lampetie, o nome Lampetusa ocorre nos manuscritos de Fulgêncio (Mit. i, 16) e nos comentários de Sérvio, sendo erroneamente atribuído a Ovídio por Boccaccio em Genealogiae (vii, 42).

Diante dessa estratigrafia de significados, a mar-

cação proposta nesta pesquisa refuta a *tag* única e simplista. Adotou-se uma anotação profunda no atributo `@ref`, vinculando o termo a múltiplos identificadores. No `<teiHeader>`, essas referências são "casadas" com as notas de Peixoto e Silva Dias, inseridas no elemento <note>. Assim, a hierarquia do XML não apenas identifica a palavra, mas documenta a erudição camoniana e a história da sua recepção crítica, transformando o arquivo TEI em uma ferramenta de interoperabilidade bibliográfica.

A literatura recente em Humanidades Digitais reforça que a codificação estruturada não é neutra, mas implica escolhas ontológicas. Ao atribuir tipos (@type) e referências (@ref), constrói-se uma camada semântica que pode ser explorada tanto para análises quantitativas quanto qualitativas. Desse modo, a edição digital deixa de ser apenas um repositório eletrônico e passa a funcionar como um laboratório interpretativo.

3 Metodologia e Desenvolvimento

Para essa pesquisa foram utilizados dois arquivos XML de base: a edição de 1572 (versão dextrógira), mantendo a ortografia arcaica e marcos codicológicos como quebras de linha (`<lb/>`) e assinaturas de cadernos (`<fw>`). E a edição modernizada do Projeto Gutenberg, uma versão normalizada com as regras ortográficas e gramaticais atuais, para facilitar a eficácia dos modelos de linguagem modernos. Ambas as versões foram submetidas ao mesmo processo de análise automática, com o objetivo de observar o comportamento das ferramentas de reconhecimento de entidades nomeadas em contextos textuais distintos, um com ortografia histórica e estrutura editorial preservada, e outro com ortografia modernizada. Essa comparação permitiu avaliar em que medida modelos contemporâneos de Processamento de Linguagem Natural conseguem lidar com textos literários clássicos em diferentes estágios de normalização. O processamento foi realizado em Python, estruturado em três etapas:

1. **Dicionário de Autoridade (Gazetteer):** Uma lista pré-definida de termos críticos garantiu a precisão de entidades frequentes.

2. **Modelo de Linguagem (NER):** Utilizou-se a biblioteca spaCy (modelo `pt_core_news_lg`) para identificar entidades não catalogadas.

3. **Algoritmo de Proteção de Estrutura:** Para a versão de 1572, desenvolveu-se um sistema de regex capaz de identificar nomes segmentados por tags de quebra de linha (ex: `Lusi<lb/>tana`), evitando a fragmentação do dado semântico.

A comparação entre os resultados obtidos nas duas versões evidenciou limitações no desempenho do modelo de linguagem para a identificação consistente das entidades nomeadas no corpus camoniano. Em razão disso, a marcação manual e a utilização do dicionário de autoridade mostraram-se estratégias mais eficazes e confiáveis para a anotação semântica do texto.

A utilização da biblioteca spaCy (modelo `pt_core_news_lg`) e de scripts em Python permitiu a extração célere de antropônimos e topônimos recorrentes. A automação garante a escalabilidade e a consistência terminológica, evitando omissões comuns em tarefas manuais exaustivas. No entanto, verificou-se um significativo *domain shift*. Modelos NER contemporâneos, treinados em dados jornalísticos, apresentam baixa performance diante da instabilidade ortográfica da edição de 1572 e da sintaxe épica. O algoritmo frequentemente falha na desambiguação contextual, classificando figuras mitológicas (ex: Marte) como locais geográficos (planeta) e ignorando epítetos complexos (ex: "o grão Macedônio"), que são tratados como substantivos comuns. Além disso, substantivos como "Fama" e "Mar" foram classificados erroneamente como pessoas devido à capitalização poética.

A intervenção manual foi aplicada para converter o texto digital em uma edição curada, funcionando como uma camada de pós-edição filológica, essencial para a resolução de casos de polissemia. Onde a Inteligência Artificial (IA) identifica apenas uma string, o pesquisador, subsidiado por Silva Dias (1913) e Afrânio Peixoto (1924), entre outros, identifica a densidade intertextual. O caso de "Lampetusa" (I, 46) é emblemático: enquanto o NER sugere um local, o editor humano codifica a referência mitológica às Helíades, utilizando o elemento `<rs>` para mapear a antonomásia. No entanto, a marcação manual é onerosa e dificilmente escalável para grandes volumes de dados, além de ser suscetível à subjetividade do anotador, o que pode gerar inconsistências estruturais sem o auxílio de esquemas de validação (como o RelaxNG).

A metodologia adotada utilizou a automação como um "primeiro estágio" de detecção estrutural e o dicionário de autoridade (Gazetteer) para fixar entidades inequívocas. O esforço humano concentrou-se no refinamento semântico de alto

nível, especificamente na tipificação de mitônimos e na estruturação de cadeias de referência. Essa simbiose entre o processamento algorítmico e a crítica textual permitiu que o XML final servisse tanto para análises quantitativas de frequência quanto para a recuperação qualitativa da tradição crítica camoniana. A hierarquia XML foi desenhada para garantir a interoperabilidade. O elemento `<teiHeader>` abriga a Prosopografia (relação de todas as pessoas mencionadas) e a Gazeta (relação de todos os lugares mencionados) do projeto, onde cada entidade possui um `xml:id` unívoco. No corpo do texto, o atributo `@ref` estabelece o vínculo semântico (*Linked Data*), permitindo a normalização de variantes (ex: "Lusitana", "Lusa" e "Portugal" convergem para `#portugal`).

4 Discussão

Os resultados da modelagem demonstram que a combinação entre métodos automáticos e curadoria humana produz um ganho significativo na representação semântica do texto épico. Enquanto o modelo NER identifica entidades explícitas com eficiência satisfatória, ele apresenta limitações na detecção de epítetos, antonomásias e formas alegóricas. A inserção de um *gazetteer* especializado mitiga parcialmente esse problema, mas não elimina a necessidade de intervenção crítica.

A estrutura TEI adotada permite consultas complexas que não seriam possíveis em uma edição linear. Por exemplo, torna-se viável recuperar todas as ocorrências de uma entidade independentemente de sua forma superficial, mapear a distribuição de mitônimos ao longo dos cantos ou analisar a frequência relativa de referências históricas e mitológicas. A presença sistemática do atributo `@ref` consolida essa interoperabilidade, aproximando a edição de princípios de *Linked Data*.

Entretanto, a formalização impõe limites. A categorização tripartida (antropônimos, mitônimos e topônimos) simplifica um universo referencial mais fluido, no qual certas figuras oscilam entre história e mito. Além disso, a tipificação actancial não foi plenamente automatizada, exigindo decisões interpretativas que podem variar conforme a tradição crítica adotada.

Apesar dessas restrições, o modelo proposto revela-se escalável e replicável para outros textos da tradição épica renascentista. A metodologia pode ser adaptada a diferentes corpora históricos, desde que acompanhada de uma etapa de revisão filológica rigorosa. O principal contributo reside na demonstração de que a edição digital, quando concebida como estrutura semântica relacional, amplia as possibilidades de investigação literária e historiográfica.

5 Considerações Finais

A metodologia aplicada neste estudo demonstrou que a marcação semântica de textos clássicos exige uma abordagem que transcenda a automação purista. Os resultados evidenciam que, embora modelos NER modernos (como o spaCy) ofereçam uma base sólida para a escalabilidade, eles são incapazes de capturar a densidade metafórica e as ambiguidades eruditas inerentes à poesia renascentista. A abordagem híbrida proposta — integrando *gazetteers* especializados e pós-edição humana — superou o isolamento algorítmico, especialmente na identificação de antonomásias e no tratamento de polissemias complexas, como o caso "Lampetusa", onde o dado geográfico e o mítico coexistem.

A principal contribuição deste trabalho para as Humanidades Digitais reside na validação do XML/TEI não apenas como um formato de arquivamento, mas como uma ferramenta de hermenêutica digital. Ao "casar" o texto poético com as autoridades lexicográficas de Afrânio Peixoto, Júlio Nogueira e Silva Dias, o corpus deixa de ser um silo de informação para tornar-se uma base de dados conectada (*Linked Data*). Esta arquitetura possibilita, em etapas futuras, a realização de análises de redes sociais (*Social Network Analysis*) para mapear a interação entre deuses e heróis, além da geração de cartografias digitais precisas das navegações lusitanas. Em última análise, a pesquisa reafirma que o futuro da Filologia Digital camoniana não reside na substituição do pesquisador pela máquina, mas na instrumentalização técnica do olhar crítico, garantindo que a erudição clássica seja preservada e potencializada no ecossistema digital.

References

Vítor Manuel de Aguiar e Silva. 2010. *Camões: Labirintos e Fascínios*. Cotovia, Lisboa.

Luís de Camões. 1916. *Os Lusíadas*. Companhia Portugueza Editora, Porto. Comentados por Augusto Epiphanio da Silva Dias. 2. ed. melhorada. Tomo I.

Algirdas Julien Greimas. 1973. *Semântica estrutural: pesquisa de método*. Cultrix, São Paulo.

Franco Moretti. 2011. Network theory, plot analysis. Pamphlet 2, Stanford Literary Lab.

Júlio Nogueira. 1960. *Dicionário e Gramática de "Os Lusíadas"*. Livraria Freitas Bastos S.A., Rio de Janeiro.

Afranio Peixoto and Pedro A. Pinto. 1924. *Dicionário d'Os Lusíadas de Luís de Camões*. Livraria Francisco Alves - Casa de Paulo Azevedo e Cia., Rio de Janeiro. Acesso em: 21 maio 2025.

TEI Consortium. 2023. Tei p5 guidelines.

A elaboração de uma edição digital d'*Os Lusíadas*

Bruno Maroneze
UFGD / Dourados, MS
brunomaroneze@ufgd.edu.br

Vanessa Martins do Monte
USP / São Paulo, SP
vmmonte@usp.br

André Bertacchi
UFJF / Juiz de Fora, MG
andre.bertacchi@ufjf.br

Artur Costrino
UFOP / Ouro Preto, MG
artur.costrino@ufop.edu.br

Alexandre Agnolon
UFOP / Ouro Preto, MG
alexandre.agnolon@ufop.edu.br

Mário Eduardo Viaro
USP / São Paulo, SP
maeviaro@usp.br

Resumo

This article presents the *Lusíadas Digital* project, which proposes the development of a virtual philological edition of *Os Lusíadas* by Luís de Camões, integrating principles of textual criticism, Digital Humanities, and Natural Language Processing (NLP). The project aims to develop a digital platform bringing together facsimiles of the 1572 editions, diplomatic and modernized transcriptions, a dynamic critical apparatus, a lexical glossary with etymological information, historical and literary commentary, and translations aligned with the original text. The methodology combines traditional philological practices with XML-TEI text encoding, OCR techniques, automatic lemmatization, version alignment, and lexical mining. Initially focused on Canto I, the project seeks to establish a scalable and replicable model for the remaining cantos of the work. By proposing an open, interoperable, and data-oriented digital infrastructure, the initiative contributes to the advancement of e-Philology in Brazil and to the development of technologies applied to the digital critical editing of manuscripts and early printed editions.

1 Introdução

A digitalização de obras clássicas deixou de ser mero processo de reprodução fac-similar e estimulou a constituição de um campo de investigação interdisciplinar, as chamadas Humanidades Digitais. A edição filológica virtual (Monte e Paixão de Sousa, 2017) representa um dos produtos mais sofisticados de integração entre saber humanístico e tecnologia computacional.

Apesar da centralidade d'*Os Lusíadas* para a tradição literária da língua portuguesa, ainda não existe uma edição filológica virtual que integre, de forma sistemática e interativa:

- imagens fac-similares das primeiras edições;
- transcrições diplomáticas e modernizadas;
- aparato crítico comparativo;
- glossário lexical estruturado;
- traduções alinhadas;
- comentários histórico-literários incorporados à navegação.

O projeto *Lusíadas Digital* busca suprir essa lacuna, propondo uma edição filológica virtual que combine rigor crítico, modelagem formal de dados e ferramentas de Processamento de Linguagem Natural (PLN).

2 Problema filológico: as edições de 1572 e a tradição textual

O problema editorial d'*Os Lusíadas* é conhecido: há duas edições datadas de 1572, tradicionalmente identificadas pela orientação do pelicano no frontispício (sinistrógira e dextrógira). As diferenças entre ambas incluem:

- variantes ortográficas;
- divergências de pontuação;
- substituições lexicais;
- lapsos tipográficos;
- diferenças na disposição gráfica.

A dificuldade em determinar a precedência entre as edições gera implicações para a crítica textual e para a história da língua portuguesa. A edição crítica de Augusto Epiphanio da Silva Dias (Camões,

Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)
Vol. 2, pages 298–302

1916) constitui referência incontornável, mas foi concebida para circulação impressa e para público especializado.

No ambiente digital, torna-se possível:

- visualizar variantes em paralelo;
- automatizar a detecção de divergências;
- representar formalmente a tradição textual;
- disponibilizar múltiplos níveis de leitura.

A edição virtual, portanto, não substitui a crítica tradicional, mas a potencializa.

3 Fundamentos teóricos: Filologia, e-Philology e modelagem textual

A Filologia tradicional sempre buscou a opção textual mais verossímil de acordo com evidências manuscritas e históricas, por meio da análise da tradição manuscrita ou impressa. Contudo, como propõem Crane et al. (Crane et al., 2008), as práticas filológicas precisam transformar-se qualitativamente diante das tecnologias digitais. A chamada e-Filologia envolve:

1. digitalização imagética de alta resolução;
2. transcrição estruturada e formalmente anotada;
3. representação explícita das variantes;
4. múltiplas formas de visualização;
5. interoperabilidade e reutilização dos dados.

A edição filológica virtual proposta ancora-se na definição de Monte e Paixão de Sousa (Monte and Paixão de Sousa, 2017), que usam o termo para se "referir ao objeto criado a partir de um trabalho que inclui a produção da réplica imagética digital do documento físico, a edição filológica digital, e as múltiplas possibilidades de representação final ou publicação digital. Todos esses processos são construídos com ferramentas e tecnologias computacionais, e são, portanto, 'digitais'; o conjunto dos processos forma o objeto que chamamos de virtual – pois que simula, representa, re-cria artificialmente os documentos originais.".

A modelagem textual em TEI-XML (TEI Consortium, 2025) constitui padrão consolidado para representar desde estruturas textuais internas (como cantos, estrofes e versos) até variantes textuais e comentários críticos.

No caso d'*Os Lusíadas*, a marcação TEI permitirá representar, entre outras informações:

- diferenças entre as versões de 1572;
- distinção entre transcrição diplomática e modernizada;
- estrutura métrica;
- alinhamento com traduções.

4 Arquitetura da edição digital

A plataforma proposta está sendo organizada em módulos integrados:

4.1 Módulo fac-similar

Este módulo possibilitará:

- Visualização de imagens (tanto da edição sinistrógira quanto da dextrógira).
- Navegação por fólio e estrofe.
- Sincronização entre imagem e transcrição.

4.2 Módulo textual

O módulo textual disponibilizará:

- A transcrição diplomática das versões de 1572;
- A transcrição modernizada (seguindo os padrões ortográficos tanto do português europeu quanto do brasileiro);
- A alternância dinâmica entre versões (dextrógira, sinistrógira, modernizada).

4.3 Aparato crítico dinâmico

Em vez de um aparato crítico "tradicional", pretende-se usar os recursos digitais para exibir:

- Visualização paralela das duas edições de 1572, bem como da versão modernizada;
- Destaque automático de variantes;
- Filtro por tipo de divergência (ortográfica, lexical, tipográfica).

4.4 Glossário interativo

O formato digital permite que o glossário seja acessado por *hyperlinks* diretamente no texto.

Inicialmente, cada palavra do texto será associada ao verbete correspondente do dicionário de Peixoto e Pinto (Peixoto and Pinto, 1924), obra de referência para o estudo do texto camoniano. Futuramente, também se pretende elaborar um glossário próprio, contendo outras informações, tais como:

- definição contextual;
- etimologia;
- frequência na obra;
- ocorrências clicáveis.

Além disso, esse formato de glossário interativo permite que se acrescentem comentários de natureza histórica, literária e analítico-interpretativa, muito necessários para a plena compreensão do texto.

4.5 Traduções alinhadas

A obra de Camões foi traduzida inúmeras vezes. Pode-se mencionar, entre outras, as traduções para o espanhol (Garces, 1591), para o latim (Faria, 1622), para o inglês (Mickle, 1776) e para o italiano (Paggi, 1658). Assim, pretende-se disponibilizar as traduções com os seguintes recursos:

- Alinhamento por estrofe ou verso, quando possível;
- Visualização lado a lado com o texto português;
- Possibilidade de comparação lexical.

Abaixo pode-se ver a imagem de uma primeira versão da tela de visualização do texto.

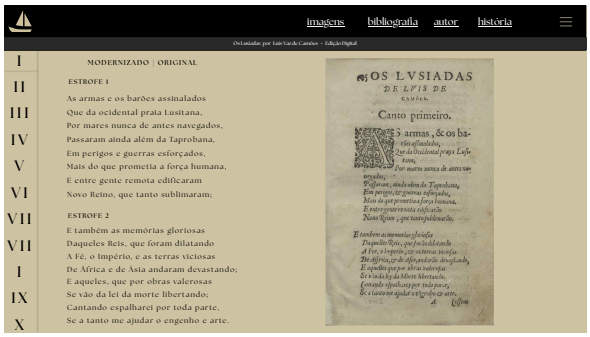


Figura 1: Primeira versão da tela de visualização do texto.

5 Metodologia

A metodologia articula etapas fundamentais do trabalho da crítica textual ao processamento computacional, consolidada em quatro frentes principais.

5.1 Levantamento e curadoria

Nesta frente, procede-se à identificação das edições fac-similares disponíveis, bem como das traduções existentes em domínio público e das obras de comentaristas e críticos (também em domínio público) que podem trazer informações relevantes a serem incorporadas na edição digital.

5.2 Digitalização e OCR

Os materiais são transcritos automaticamente por meio do OCR (*Optical Character Recognition*) na ferramenta Transkribus. Após a transcrição, é feita uma revisão manual. A transcrição do Canto I já foi inteiramente finalizada.

Como "subproduto"deste projeto, será criado um modelo de OCR para o Transkribus, treinado especificamente para a tipografia portuguesa do século XVI.

5.3 Estruturação em TEI-XML

O texto transcrito é convertido para o formato estruturado TEI-XML, com a marcação de elementos XML tais como:

- versos (`<l>`);
- estrofes (`<lg>`);
- variantes (`<app>`, `<rdg>`);
- notas (`<note>`);
- nomes próprios (`<name>`).

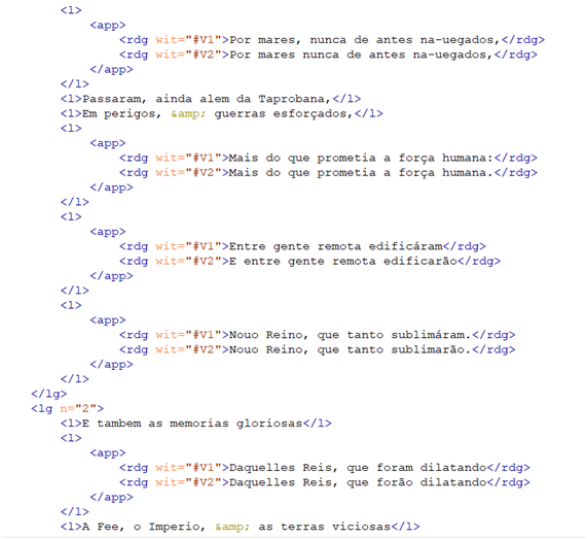
Na figura 2 a seguir (captura de tela do *software* Oxygen XML), podem-se observar as diferenças entre as versões do texto, marcadas com o elemento `<rdg>`.

5.4 Processamento de Linguagem Natural

É necessário avaliar de que forma técnicas de PLN podem ser empregadas no texto camoniano. É possível pensar em três frentes principais:

a) Lematização e etiquetagem morfossintática

A lematização é necessária para que sejam feitos cálculos de frequência lexical, bem como para a associação de cada palavra ao verbete



```
    <l>
        <app>
            <rdg wit="#V1">Por mares, nunca de antes na-uegados,</rdg>
            <rdg wit="#V2">Por mares nunca de antes na-uegados,</rdg>
        </app>
    </l>
    <l>Passaram, ainda alem da Taprobana,</l>
    <l>Em perigos, &amp;amp; guerras esforçados,</l>
    <l>
        <app>
            <rdg wit="#V1">Mais do que prometia a força humana:</rdg>
            <rdg wit="#V2">Mais do que prometia a força humana.</rdg>
        </app>
    </l>
    <l>
        <app>
            <rdg wit="#V1">Entre gente remota edificáram</rdg>
            <rdg wit="#V2">E entre gente remota edificarão</rdg>
        </app>
    </l>
    <l>
        <app>
            <rdg wit="#V1">Nouo Reino, que tanto sublimáram.</rdg>
            <rdg wit="#V2">Nouo Reino, que tanto sublimarão.</rdg>
        </app>
    </l>
</lg>
<lg n="2">
    <l>E tambem as memorias gloriosas</l>
    <l>
        <app>
            <rdg wit="#V1">Daquelles Reis, que foram dilatando</rdg>
            <rdg wit="#V2">Daquelles Reis, que forão dilatando</rdg>
        </app>
    </l>
    <l>A Fee, o Imperio, &amp;amp; as terras viciosas</l>
```

Figura 2: Trecho do texto com as marcações TEI-XML.

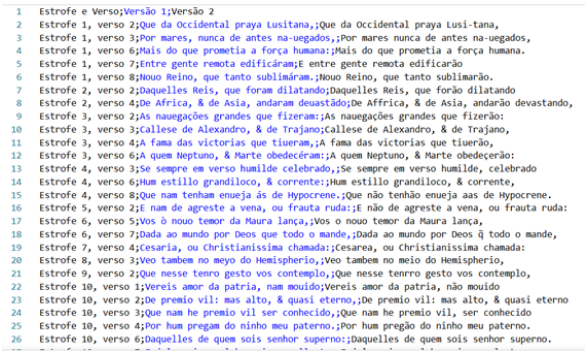
correspondente. Já a etiquetagem morfossintática possibilita buscas por construções sintáticas, o que é relevante para os estudos da linguagem literária.

Ferramentas de PLN usadas normalmente para lematização e etiquetagem (como o pacote spaCy) são treinadas com base em textos contemporâneos. Para o caso de textos do século XVI, é preciso adaptá-las para a multiplicidade de grafias da época (por exemplo, *Africa* e *Affrica*, *capitam* e *capitão*), ou fazer uma revisão "manual". Além disso, o texto poético, por ter sintaxe distinta do texto em prosa, traz mais dificuldade na etiquetagem morfossintática.

b) Detecção automática de variantes

O mapeamento das diferenças entre as edições pode ser feito por meio de algoritmos de alinhamento textual. Uma possibilidade simples (mas limitada) é o alinhamento verso a verso. A comparação automática verso a verso das duas versões do Canto I (sinistrógira e dextrógira), efetuada por meio de um algoritmo simples de comparação escrito em Python (conforme a figura 3, a seguir) detectou 238 versos divergentes (de um total de 848 versos).

No entanto, como pode haver múltiplas divergências em cada verso, o alinhamento precisa ser feito palavra por palavra, considerando também os sinais de pontuação. Uma complicação adicional surge nos casos em que



```
Estrofe e Verso;Versão 1;Versão 2
Estrofe 1, verso 2;Que da Occidental praya Lusitana,;Que da Occidental praya Lusi-tana,
Estrofe 1, verso 3;Por mares, nunca de antes na-uegados,;Por mares nunca de antes na-uegados,
Estrofe 1, verso 6;Mais do que prometia a força humana:;Mais do que prometia a força humana.
Estrofe 1, verso 7;Entre gente remota edificáram;E entre gente remota edificarão
Estrofe 1, verso 8;Nouo Reino, que tanto sublimáram.;Nouo Reino, que tanto sublimarão.
Estrofe 2, verso 2;Daquelles Reis, que foram dilatando;Daquelles Reis, que forão dilatando
Estrofe 2, verso 4;De Africa, & de Asia, andaram deuastão;De Affrica, & de Asia, andarão devastando,
Estrofe 3, verso 2;As nauegações grandes que fizeram:;As nauegações grandes que fizerão:
Estrofe 3, verso 3;Callese de Alexandro, & de Trajano;Callese de Alexandro, & de Trajano,
Estrofe 3, verso 4;A fama das victorias que tiueram,;A fama das victorias que tiuerão,
Estrofe 3, verso 6;A quem Neptuno, & Marte obedecéram:;A quem Neptuno, & Marte obedeçerão:
Estrofe 4, verso 3;Se sempre em verso humilde celebrado,;Se sempre em verso humilde, celebrado
Estrofe 4, verso 6;Hum estillo grandiloco, & corrente:;Hum estillo grandiloco, & corrente,
Estrofe 4, verso 8;Que nam tenham enueja ás de Hypocrene.;Que não tenhão enueja aas de Hypocrene.
Estrofe 5, verso 2;E nam de agreste a vena, ou frauta ruda:;E não de agreste a vena, ou frauta ruda:
Estrofe 6, verso 5;Vos ò nouo temor da Maura lança,;Vos o nouo temor da Maura lança,
Estrofe 6, verso 7;Dada ao mundo por Deos que todo o mande,;Dada ao mundo por Deos q todo o mande,
Estrofe 7, verso 4;Cesaria, ou Christianissima chamada:;Cesarea, ou Christianissima chamada:
Estrofe 8, verso 3;Veo tambem no meyo do Hemispherio,;Veo tambem no meio do Hemispherio,
Estrofe 9, verso 2;Que nesse tenro gesto vos contemplo,;Que nesse tenrro gesto vos contemplo,
Estrofe 10, verso 1;Vereis amor da patria, nam mouido;Vereis amor da patria, não mouido
Estrofe 10, verso 2;De premio vil: mas alto, & quasi eterno,;De premio vil: mas alto, & quasi eterno
Estrofe 10, verso 3;Que nam he premio vil ser conhecido,;Que nam he premio vil, ser conhecido
Estrofe 10, verso 4;Por hum pregam do ninho meu paterno.;Por hum pregão do ninho meu paterno.
Estrofe 10, verso 6;Daquelles de quem sois senhor superno:;Daquelles de quem sois senhor superno.
```

Figura 3: Listagem de diferenças entre as versões, gerada automaticamente.

palavras inteiras encontram-se faltantes numa das versões, como o artigo "o"no exemplo a seguir:

> *Eis nos bateis fogo se leuanta,* (versão dextrógira)
> *Eis nos bateis o fogo se leuanta,* (versão sinistrógira)

(Canto I, estrofe 89, verso 1)

Assim, faz-se necessário o emprego de algoritmos mais complexos, que estão sendo avaliados.

c) Alinhamento com traduções

O alinhamento do texto português com as traduções exige também o emprego de técnicas avançadas, visto que os textos traduzidos não correspondem ao texto original verso a verso. Assim, é possível que cada tradução precise ser alinhada de forma específica.

6 Desafios computacionais e inovações tecnológicas

A aplicação de PLN a textos da Primeira Modernidade apresenta alguns desafios, em especial a ausência de padronização ortográfica da época; tendo em vista que as ferramentas disponíveis são treinadas e programadas para o português atual (séculos XX e XXI), faz-se necessário recorrer a diversas adaptações.

Dessa forma, espera-se que o projeto contribua também para a criação de recursos computacionais que possam ser aplicados em outros textos históricos, tais como modelos de lematização e dicionários de formas ortográficas.

O projeto inova no sentido de que propõe:

1. Integração orgânica entre edição crítica e PLN;

2. Representação estruturada da tradição textual;

3. Interface interativa voltada tanto para especialistas quanto para estudantes;

4. Produção de dados reutilizáveis em pesquisa linguística.

Diferentemente de repositórios como o Project Gutenberg, que oferecem texto linearizado, o projeto propõe uma arquitetura baseada em dados estruturados e interoperáveis.

7 Resultados esperados e impactos

O principal resultado esperado com o projeto é a própria edição crítica digital do Canto I, com os vários recursos de consulta pretendidos. Parte integrante dessa edição crítica é o próprio texto anotado em TEI-XML, que poderá ser reutilizado em diversos outros projetos. Além disso, os algoritmos desenvolvidos poderão ser empregados em projetos similares, bem como o modelo de transcrição gerado pelo Transkribus poderá ser reaproveitado para outros textos da época. Por fim, também serão publicados artigos científicos sobre a metodologia empregada.

A médio prazo, o modelo será expandido aos demais cantos e a outras obras do período.

O projeto visa trazer impactos para:

7.1 Pesquisa filológica

Apesar de o texto camoniano já ter sido largamente estudado, a abordagem computacional pode trazer um olhar novo à obra (ainda que não necessariamente superior). Além disso, os métodos desenvolvidos podem ser aplicados a outras obras no futuro.

7.2 Linguística histórica

O texto d'*Os Lusíadas* anotado e estruturado computacionalmente servirá para apoiar estudos diacrônicos da língua portuguesa.

7.3 Ensino

Por ser *Os Lusíadas* um dos textos literários mais importantes da língua portuguesa, é uma obra amplamente estudada tanto na educação básica quanto no ensino superior. Assim, o projeto tem potencial para se tornar um recurso pedagógico interativo de grande relevância.

7.4 Desenvolvimento tecnológico

O principal impacto pretendido quanto ao desenvolvimento tecnológico é a reutilização dos algoritmos utilizados para outros projetos futuros, estabelecendo, assim, uma proposta de modelo de edição crítica digital.

8 Conclusão

O projeto Lusíadas Digital representa uma convergência entre tradição filológica e inovação tecnológica. Ao integrar modelagem textual, crítica editorial e processamento de linguagem natural, propõe-se um novo paradigma para a edição digital de textos clássicos em língua portuguesa. A edição do Canto I funcionará como projeto-piloto para uma iniciativa de maior alcance, capaz de posicionar a pesquisa brasileira em Humanidades Digitais em diálogo direto com experiências internacionais consolidadas.

References

Luís Vaz de Camões. 1916. *Os Lusíadas*. Companhia Portugueza Editora, Porto. Comentados por Augusto Epiphanio da Silva Dias. 2. ed. melhorada. Tomo I.

Gregory Crane, David Bamman, and Alison Jones. 2008. ephilology: When the books talk to their readers. In Ray Siemens and Susan Schreibman, editors, *A Companion to Digital Literary Studies*. Blackwell, Oxford.

Tomás de Faria. 1622. *Lusiadum Libri Decem*. Ex Officina Gerardi de Vinca, Lisboa.

Henrique Garces. 1591. *Los Lusiadas de Luys de Camoes*. Impresso con licencia en casa de Guillermo Dreuy, Madri. Translator.

William Julius Mickle. 1776. *The Lusiad; or, the Discovery of India. An Epic Poem*. Printed by Jackson and Lister, Oxford. Translator.

Vanessa Martins do Monte and Maria Clara Paixão de Sousa. 2017. Por uma filologia virtual: o caso das atas da câmara de são paulo (1562–1596). *Revista da Abralin*, 16:239–264.

Carlo Antonio Paggi. 1658. *Lusiada Italiana. Poema Heroico del grande Luigi de Camões*. por Henrique Valente de Oliveira, Lisboa.

Afrânio Peixoto and Pedro A. Pinto. 1924. *Dicionario d'Os Lusíadas de Luis de Camões*. Francisco Alves, Rio de Janeiro.

TEI Consortium. 2025. TEI: Guidelines for Electronic Text Encoding and Interchange. Accessed: 18 Feb. 2026.

The F1 of Formula One: Applicability of Pre-trained NER Models to Brazilian TV Interview Transcripts

João Pedro Gonçalves Munhoz[1], Luiz Felipe Guidorizzi de Oliveira[2], Isabella Belchior[1], Evandro Eduardo Seron Ruiz[2] e Oto Araújo Vale[1]
[1]Departamento de Letras, Universidade Federal de São Carlos (UFSCar)
13365-905 São Carlos, SP – Brasil,
[2]Departamento de Computação e Matemática, Universidade de São Paulo (USP)
14040-901 Ribeirão Preto, SP – Brasil
Correspondence: otovale@ufscar.br

Abstract

Recorded interviews can capture their subjects' memories, perceptions, and emotions. When conducted with notable figures, they also have the potential to serve as a resource for interdisciplinary research, impacting various branches of science. In this work, we mark the beginning of a significant project analyzing interviews from the Roda Viva program, the longest-running interview show on Brazilian television. In this initial study, we examined six memorable interviews with six Brazilian Formula One drivers to compare the performance of two named entity recognition methods: a statistical-neural method and large language models, both evaluated against manual annotations. Still, it highlighted relevant qualitative distinctions: the statistical method showed a rigid dependence on capitalisation and lexical familiarity, leading to mechanical false positives and missing non-capitalised entities, while the LLM exhibited greater linguistic sensitivity, retrieving contextual entities and being robust to transcription errors, though it still produces false positives. The LLM-based model appears more promising due to its flexibility and the potential for refinement via instructions to filter for ambiguities, favouring the automation of social network extraction in the corpus.

1 Introduction

Television interviews have been increasingly consolidated as fundamental sources for research in digital humanities, offering a rich multimodal repository that combines verbal language, body expression, intonation, and visual context, elements essential for reconstructing historical narratives and mapping social networks over time. Unlike static textual documents, these audiovisual records capture not only what was said, but also how, when, and by whom, enabling denser analyses of collective memory construction, the formation of national imaginaries, and the dynamics of power relations within specific historical contexts. In this landscape, long-form interview programs such as *Roda Viva*[1] assume particular relevance, as they constitute continuous documentary series that have recorded, over decades, the voices of leading figures in Brazilian politics, culture, science, and sports.

The current Roda Viva corpus comprises 713 transcribed interviews available on the Memória Roda Viva portal[2], compiled into machine-readable formats by de Miranda Jr et al., 2024. This corpus documents more than three decades of contemporary Brazilian history, forming a documentary repository of inestimable value for understanding the formation of our collective memory.

Ideally, comprehensive processing of this corpus would allow, among other things, the construction of an extensive named entity network—a digital map of the personal, political, and cultural connections woven across thousands of hours of public dialogue. However, before undertaking analysis at such scale, rigorous methodological validation of entity extraction and identification techniques is required.

Although transcribed interviews do not faithfully reproduce what occurred during the televised interviews, they can be considered a genre in their own right. They do not fully conform to traditional written texts, since the sequence of utterances (marked by interruptions and frequent turn-taking) differs significantly from the patterns found in news journalism or opinion pieces. On the other hand, it is evident that there are multiple levels of orality representation. While hesitations and repetitions are generally omitted to ensure readability, the flow of information retains the essential characteristic of this type of Roda Viva interview format: that of an interviewee responding alternately to multiple interviewers, thereby conferring upon the text a

[1]https://cultura.uol.com.br/programas/rodaviva/
[2]https://rodaviva.fapesp.br/

Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)
Vol. 2, pages 303–311
April 13-16, 2026

dialogic and fragmented structure.

In this paper, we present a pilot study that takes as its test case a cohesive and culturally significant subset of the corpus: interviews conducted with Brazilian Formula One drivers. This selection is justified not only by the controlled yet relationally complex environment it offers for evaluating named entity recognition techniques, but also by the domain's symbolic relevance. Interviews in this subset contain dense networks of references to on-track rivals, team principals, engineers, sponsors, specialized journalists, and family members—relationships that are historically documented and amenable to systematic computational analysis.

Formula One occupies a singular place in the national imaginary: more than an elite sport, it has become a stage for projecting Brazilian identity on the global scene, with its drivers assuming roles as cultural ambassadors and, in emblematic cases, national heroes. Within this context, Ayrton Senna transcended the racetrack to become a symbol of excellence, determination, and patriotism: a contemporary myth whose trajectory, prematurely cut short, continues to shape narratives about Brazil and its place in the world. Senna's centrality as a catalyst of memories and social connections makes this subset especially fertile for investigating how public narratives construct and preserve symbolic bonds among individuals, institutions, and the nation.

This subcorpus includes archived interviews with Ayrton Senna (1986), Nelson Piquet (1994), Emerson Fittipaldi (1995), and Rubens Barrichello (1996), supplemented by our own transcriptions of interviews with Christian Fittipaldi (1995) and Lucas di Grassi (2022). By focusing on this specific domain, we aim to establish and validate a robust pipeline for Proper Name Recognition. This approach ensures that the extraction criteria remain reliable and transparent, providing a consistent framework for the future expansion of the study to the full corpus. The results discussed here not only contribute to the preservation and critical analysis of Brazilian sports memory, highlighting the legacy of its greatest hero, but also pave the methodological path for large-scale relational network construction from multimodal audiovisual sources.

The remainder of this paper is organized as follows: Section 2 reviews related works to situate our study within the current literature. Section 3 details the data and methods employed in our analysis, followed by a comprehensive presentation and discussion of our results in Section 4. The paper concludes in Section 5 with a summary of findings.

2 Related work

Automated extraction of content and relationships between those contents from interviews has been little explored, according to our literature review. Husevåg, 2019 investigates the potential of subtitles as a source for automatic indexing of TV programs through named entity recognition (NER), finding that while subtitles capture a substantial subset of salient entities, especially personal names across genres and creative works in literature programs, they alone cannot fully replicate manually created metadata. In Adriansen, 2012, the authors address one of the first studies of interviews as material of historical and social interest, explaining how researchers can utilize interviews as a tool for conducting life history research. We also found very few academic articles on enhancing historical knowledge from interviews, and the automated extraction of social relations described in these interviews remains a relatively unexplored topic, as noted by (Laato et al., 2025). Laato and his team conducted a zero-shot information-extraction study on 89,339 brief Finnish interviews with refugee families relocated after WWII. They extract social organizations and hobbies for each family member as proxies for social integration, and compare several generative models using a supervised approach to evaluate their relative strengths.

In another study, Hicke et al., 2025 has shown that researchers can expand their understanding of history and society with the help of Natural Language Processing resources and large language models. In their article, Hicke and co-workers adapted human-annotated prompts for large language models to identify and characterize portrayals of acts of God in a corpus of 88 Christian fiction novels. Similarly, (Poibeau, 2024) assessed large language models for annotating Roman and Greek mythological references in modern French literature, presented an annotation scheme, and showed how LLMs can follow it effectively despite occasional significant errors. His study includes graphically relating people, organizations, and other entities mentioned in these interviews.

Researchers recognize that the potential social relations between named individuals in the interviews involve the computational task of Named Entity

Recognition (NER). This task identifies mentions of rigid designators in free text related to predefined semantic types, such as persons, places, and organizations. A rigid designator is defined as a term that identifies the same entity across all 'possible worlds' in which that entity exists (Kripke, 1972). For instance, the name 'Aristotle' functions as a rigid designator because it refers to a specific individual regardless of the counterfactual circumstances or descriptions associated with him. Li et al., 2020 have published a valuable survey on NER.

The field of Named Entity Recognition (NER) in Portuguese has matured significantly since the inaugural HAREM evaluation contests (Santos et al., 2006; Freitas et al., 2010), progressing toward contemporary benchmarks that test the efficacy of Large Language Models (LLMs) in specialized domains. However, while we acknowledge the significant initiatives advancing the state-of-the-art—most notably the pre-trained BERTimbau transformer (Souza et al., 2020) and similar architectures (Souza et al., 2023) applied in sectors ranging from healthcare (Schneider et al., 2020) to jurisprudence (Nunes et al., 2024) – research remains predominantly focused on conventional written corpora. These genres are typically characterized by an objective or declarative tone, which contrasts sharply with the dialogic and spontaneous nature of transcribed interviews.

Consequently, the primary objective of this study is to evaluate standard off-the-shelf approaches applied to real-time, semi-spontaneous conversations, specifically television interviews. Consequently, we focus our evaluation on two distinct architectures: a dedicated neural model from an industry-standard NLP library and an open-weight Large Language Model (LLM).

3 Data & Methods

Data

The program Roda Viva is one of the longest-running interview programs on Brazilian television[1]. It has aired every Monday at 22:00 on TV Cultura since 1986. You can freely access all 27 seasons of interviews on the program's YouTube channel[2]. Researchers, students, viewers, and internet users can explore 713 transcribed interviews (de Miranda Jr et al., 2024)[3], which provide content in text form, complete with entries, references, photographs, and short videos.

In this initial project, we compare named-entity retrieval methods for extracting person mentions from Roda Viva interviews, evaluating a statistical approach, large language models, and manual annotation to see how effectively each captures the social links described in free text.

We selected interviews with six Brazilian Formula One drivers:

1. Ayrton Senna da Silva. Three-time Formula One World Champion (1988, 1990, 1991). Interviewed in 1986;

2. Nelson Piquet. Three-time Formula One World Champion, who competed in 204 races from 1978 to 1991, with 23 victories during that period. Interviewed in 1994;

3. Rubens Barrichello. Held the record for the longest uninterrupted participation in the Formula One World Championship from 1993 to 2011. Interviewed in 1995;

4. Christian Fittipaldi. Participated in 43 Formula One races between 1992 and 1994. Interviewed in 1995; and

5. Emerson Fittipaldi. Two-time Formula One World Champion (1972, 1974). Interviewed in 1995.

6. Lucas di Grassi. Participated in 18 Formula One races in 2010. 2016 FIA[4] Endurance Vice-Champion. Interviewed in 2022.

Table 1 presents descriptive statistics regarding the distribution of entities within the interviews.

Interview	Unique entities	Number of sentences	Entity density
Senna	66	208	122.59
Piquet	81	309	91.24
Barrichello	103	277	80.22
Christian	73	243	115.17
Emerson	124	269	60.15
Di Grassi	73	214	119.50

Table 1: Descriptive statistics of the dataset, showing the number of unique entities, number of sentences, and entity density.

[3] https://github.com/LeGOS-UFSCar/Roda-Viva/tree/main/Corpus/V0-2/csv

[4] Fédération Internationale de l'Automobile.

Methods

We retrieved the annotated text of the six interviews. Table 2 lists the total number of tokens in these interviews.

For all interviews, we extracted the names mentioned by the interviewee in three different ways:

1. Manual annotation. Four linguists from the Department of Letters at the Federal University of São Carlos (UFSCar) manually annotated all six interviews. Since two interviews were not included in the Projeto Memória Roda Viva[5], the linguists also performed the manual transcriptions before annotating them.

2. Neural statistical method. We used a neural transition-based model with a convolutional feature extractor (CNN) and residual connections for named entity recognition, as described in (Honnibal and Montani, 2017). We implemented this using the spaCy module and loaded the auxiliary model `pt_core_news_lg`.

3. Large language model (LLM) prompt. For this project, we adopted Ollama's `gpt-oss:20B`[6], a recent open-weight model that Ollama[7] and OpenAI[8] developed in partnership.

4 Results

Table 2 shows the number of tokens per interview.

Interview	# of tokens
Senna	15,814
Piquet	23267
Barrichello	22,863
Christian	22,688
Emerson	23,278
Di Grassi	26,050

Table 2: Number of tokens per interview.

We annotated all six interviews using the three methods described above, focusing solely on identifying personal names. We consider manual annotation the gold standard and validate the other annotations against it.

Although relatively normalized during the transcription process, these interviews, as records of oral speech, exhibit marks of orality and a degree of spontaneity characteristic of the Roda Viva program. The interviewees are seated in the center, with the interviewers arranged in a semicircle around them. In every interview, there is a moderator whose role is to facilitate communication and to ensure an impartial, balanced, and productive process. While the program's tradition confers a formal character on the interviews, the spontaneity of the dialogue means that name designations are not always uniform. Let us consider, for example, some of the personal names used in the interview with former driver Ayrton Senna: 'Ayrton Senna da Silva', his full name, mentioned only once; 'Senna', once; 'Ayrton', fifteen times; 'Ayrton Senna', seven times. In other words, there are several denominations that all refer to the same person. This variability of expression does not occur only with the interviewee, that is, the person with the longest speaking time in the interview, but also with other named individuals, such as 'Lauda' and 'Nick Lauda'[9], a three-time Formula One World Champion and, for two seasons, in 32 Grands Prix, Senna's opponent. Similar cases occur in the other four interviews and also in the recognition method based on an LLM. As a point of interest, still in the case of Ayrton Senna's interview, while the neural statistical method recognized 57 named individuals, the LLM recognized 74. The gold standard identified 66 individuals.

Although multiple denominations for the same person occur, we compared the names identified by neural statistical and LLM-based methods with the gold standard at the character level. In this approach, we considered two names identical if their characters matched.

Evaluation

To evaluate the performance of each method, we report precision, recall, and the F1-score, where precision measures the proportion of retrieved items that are relevant, recall measures the proportion of relevant items that are successfully retrieved, and the F1-score is the harmonic mean of precision and recall, providing a single summary indicator that balances both dimensions. Formally, these metrics are given by:

[5] https://rodaviva.fapesp.br/materia/207/roda_viva/sobre_o_projeto.htm

[6] https://ollama.com/library/gpt-oss

[7] https://ollama.com/

[8] https://openai.com/

[9] Sic. The entity refers to Niki Lauda; the form "Nick" appears in the source transcription.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$F_1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

Where:

- TP (True Positives): are the correct positive predictions;
- FP (False Positives): are the predicted positive but actually negative; and,
- FN (False Negatives): stands for the predicted negative but actually positive

Table 3 shows the TP and FP counts for both methods, the statistical neural method, and the LLM. It also shows the counts for names not found.

Interview	TP	FP	Not Found
Senna	52	5	14
Senna (LLM)	62	13	4
Piquet	62	23	19
Piquet (LLM)	68	16	13
Barrichello	75	18	28
Barrichello (LLM)	93	21	10
Christian	51	23	22
Christian (LLM)	64	33	9
Emerson	100	19	24
Emerson (LLM)	102	11	22
Di Grassi	56	28	17
Di Grassi (LLM)	64	15	9

Table 3: Counts for true positives, false positives, and names not found for the statistical neural method and the LLM.

Next, Table 5 reports the precision, recall, and F1-score of the neural statistical model for each interview. Overall, the neural model achieves acceptable recall across all interviews, indicating that it retrieves most of the names in the gold standard. However, precision varies substantially across interviews: while it is strong for Senna (0.91) and moderately high for Emerson (0.84) and Barrichello (0.81), it is lower for Piquet (0.76) and Christian (0.70), suggesting a higher rate of false positives in these cases.

The confusion matrices generated through the application of the neural statistical method, presented in Table 4, substantiate this condition. This discrepancy results in F1-scores that are reasonably balanced only for Senna (0.84) and Emerson (0.82), whereas Piquet, Barrichello, Di Grassi and Christian show considerably weaker overall performance (0.75, 0.76, 0.71 and 0.69, respectively).

Table 4: Confusion Matrices under the statistical neural method

		Predicted	
		Positive	**Negative**
Actual	**Pos**	62 (TP)	19 (FN)
	Neg	23 (FP)	0

(a) Performance for the Piquet interview.

		Predicted	
		Positive	**Negative**
Actual	**Pos**	51 (TP)	22 (FN)
	Neg	23 (FP)	0

(b) Performance for the Christian interview.

In the Senna interview, the model achieves an F1-score of 0.84, the highest among all cases for this method. The Table 5, therefore, reveals that the model is not uniformly robust across interviews and may be sensitive to interview-specific characteristics, such as lexical variation, discourse structure, or annotation idiosyncrasies.

Interview	Precision	Recall	F1-score
Senna	0.91	0.79	0.84
Piquet	0.73	0.76	0.75
Barrichello	0.81	0.73	0.76
Christian	0.69	0.70	0.69
Emerson	0.84	0.81	0.82
Di Grassi	0.67	0.77	0.71

Table 5: Evaluation metrics for the neural statistical model.

Similarly, Table 7 presents the precision, recall, and F1-score of the LLM-based method for each interview. Compared to the neural statistical model, the LLM achieves systematically higher recall, particularly for the Senna, Barrichello, and Christian interviews (all with recall greater than or equal to

0.88), indicating that it retrieves a larger proportion of the names present in the gold standard. However, this gain in recall may come at the expense of precision, which remains very low for Christian (0.66), revealing a substantial number of false positives and suggesting a tendency to over-generate named entities. The confusion matrices obtained from the implementation of the LLM method, as illustrated in Table 6, provide compelling evidence for this condition. As a result, the F1-scores for the Christian interview (0.75) remain modest, and all the others show clearly reasonable overall performance. We also see that, in the case of Senna, both recall and precision are high. Thus, while the LLM model is effective in not "missing" names, it lacks consistent reliability across interviews and appears particularly prone to spurious name recognition in a single case.

		Predicted	
		Positive	**Negative**
Actual	**Positive**	64 (TP)	9 (FN)
	Negative	33 (FP)	TN

Table 6: Confusion Matrix for Christian under the LLM method.

Interview	**Precision**	**Recall**	**F1-score**
Senna	0.83	0.94	0.88
Piquet	0.81	0.84	0.82
Barrichello	0.82	0.90	0.86
Christian	0.66	0.88	0.75
Emerson	0.90	0.82	0.86
Di Grassi	0.81	0.88	0.84

Table 7: Evaluation metrics for the LLM model.

A rough comparison of these two approaches highlights differences among the three metrics. In Table 8, we display the percentile differences between the valuation metrics for the LLM model and the statistical neural model.

With the exception of precision in the Senna and Christian interviews, the Ollama LLM outperformed the statistical neural model.

Interview	**Precision**	**Recall**	**F1-score**
Senna	-0.08	0.15	0.04
Piquet	0.08	0.08	0.07
Barrichello	0.01	0.17	0.10
Christian	-0.03	0.18	0.06
Emerson	0.06	0.01	0.04
Di Grassi	0.14	0.11	0.13

Table 8: Estimating the difference in percentile between the valuation metrics for the LLM model and the statistical neural model.

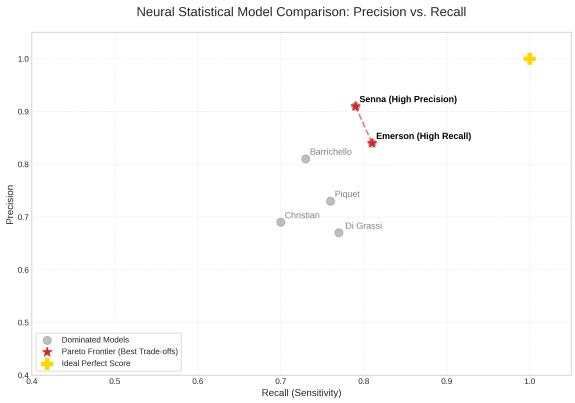


Figure 1: Comparison of precision X recall for the neural statistical model.

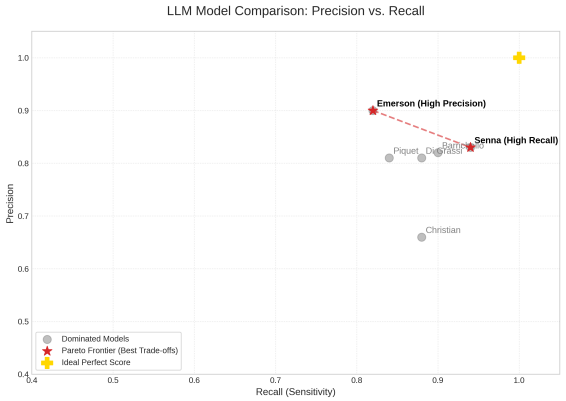


Figure 2: Comparison of precision X recall for the LLM model.

Figures 1 and 2 above illustrate the precision-recall performance for both models. Additionally, we have drawn a Pareto line (or Pareto Frontier) to delineate the boundary of optimal performance. This frontier connects the set of 'non-dominated' configurations, implying that no other option offers a superior combination of both metrics. Consequently, points lying on this line represent the most efficient trade-offs available.

5 Conclusion

This study evaluated two distinct approaches for retrieving named entities from the Roda Viva interview archive: a neural statistical method and a large language model. Quantitatively, the difference between the two methods was not statistically significant as initially expected, with p-values consistently exceeding the 0.05 threshold across all six interviews analyzed. However, a qualitative analysis of errors and 'false positives' reveals a fundamental divergence in how each model interprets the concept of a named entity in Brazilian Portuguese.

The neural statistical method demonstrated a rigid reliance on lexical familiarity and orthographic features, specifically capitalization. While this aligns with standard Portuguese grammar, where proper nouns use title casing, the neural model struggled to differentiate between named entities and simple sentence initiators, incorrectly tagging phrases such as *'É. Exato'* and *'É. Minha'* solely due to their casing. Furthermore, the neural statistical model appeared to treat lexical anomalies as probable proper nouns. Tokens such as *'eh'* and *'Ã'* (likely unintelligible to the model lexically) were frequently tagged as entities, suggesting that the model categorizes unknown or out-of-vocabulary terms as proper names by default. Conversely, the model failed to identify valid entities that lacked standard capitalization, resulting in lower recall in complex scenarios than the generative approach.

In contrast, the LLM demonstrated greater productivity and linguistic sensitivity, showing less sensitivity to capitalization and a more semantic focus. It successfully retrieved contextually correct entities that human annotators missed, such as *'a mãe do Rubinho'* (Rubinho's mother) and *'tio Emerson'* (Uncle Emerson). Notably, the LLM demonstrated robustness to transcription errors: it correctly identified *'Ayrotn'* (a typo introduced by transcribers from the previous project) as a named entity, whereas the neural statistical method failed to detect it. While it remains to be seen if the LLM can successfully link this orthographically deviant form to the specific 'Ayrton Senna' entity in a downstream resolution task, this detection capability highlights the model's semantic focus over strict orthographic matching. This is highly valuable for Digital Humanities, where transcriptions can suffer from archival inconsistencies or transcription errors such as 'Ayrotn'. The capacity of LLMs to prioritize semantic context over rigid character matching makes them a more reliable resource for the unsupervised, *en masse* processing of unlabeled historical text, ensuring that actors within the archive remain visible even when the digital record is imperfect.

Interestingly, both models showed similar limitations in entity boundary assignment in possessive constructions. In instances such as *'meu filho Luca'* (my son Luca) and *'minha filha Joana'* (my daughter Joana), both the statistical method and the LLM extracted only the proper names (*'Luca'*, *'Joana'*) rather than the complete descriptive noun phrase. This suggests that, without a defined framework explicitly instructing the methods to identify the longest possible entity span, both approaches tend to default to the specific proper noun rather than the relational context.

While the LLM's sensitivity led to specific types of false positives that differed significantly from the mechanical errors of the statistical model, these results offer a unique benefit for downstream processing. The LLM frequently annotated highly deictic expressions such as *'seu pai'* (your father), *'tua irmã'* (your sister), and *'meu pai'* (my father), as well as personal pronouns. Although these terms semantically refer to persons, they were excluded from the manual gold standard because effectively incorporating them would require an additional layer of annotation focused on resolving relations between entities, i.e. coreference resolution (Liu et al., 2023).

However, rather than viewing these as simple errors, we argue that this sensitivity is a methodological advantage. In a Digital Humanities context, capturing these deictic markers is a crucial first step for entity linking and social graph extraction. If a future model is capable of correctly performing entity linking, these "false positives" become high-value nodes that link individuals through kinship and social proximity, providing a much denser map of the Roda Viva archive than proper nouns alone. Without this initial capture, such deep relational information could remain invisible to unsupervised, *en masse* processing.

It is important to note that these localized, personal references may offer diminishing returns for the project's purpose. Because these entities are often unique to a single interview's narrative, they risk remaining as isolated nodes. Unlike public figures who appear across decades of the Roda

Viva corpus, these specific relations may not contribute to the 'global scheme' of the social network, ultimately offering limited value for large-scale, cross-interview relational mapping.

Ultimately, while both methods achieved comparable F1-scores, the LLM shows greater promise for future iterations of this project. Its false positives are linguistically grounded rather than orthographically accidental, making them methodologically more manageable. The generative nature of the model allows for the implementation of improved system instructions (such as negative constraints to ignore pronouns or a stricter reference framework) to filter out these ambiguities. Therefore, despite current statistical parity, the LLM offers a more flexible and robust approach to automating the extraction of social networks from the Roda Viva corpus.

This study validates the Roda Viva corpus not only as an audiovisual archive but also as a powerful textual resource for Digital Humanities, essential for the reconstruction of historical narratives and the dynamic mapping of social networks. By demonstrating that large language models (LLMs) offer the necessary semantic flexibility to process the spontaneity of televised speech and withstand transcription errors, this study overcomes the initial methodological barrier to the large-scale processing of this corpus. Thus, the findings presented here pave the way for transforming thousands of hours of public discourse into a structured digital map of the political, cultural, and personal connections that weave the Brazilian collective memory, thereby fulfilling the purpose of rendering visible the power dynamics and national imaginaries preserved across these decades of interviews.

Acknowledgement

This work was carried out at the Center for Artificial Intelligence of the University of São Paulo (C4AI – http://c4ai.inova.usp.br/), with support by the São Paulo Research Foundation (FAPESP Grant #2019/07665-4) and by the IBM Corporation.

References

Hanne Kirstine Adriansen. 2012. Timeline interviews: A tool for conducting life history research. *Qualitative Studies*, 3(1):40–55.

Isaac Souza de Miranda Jr, Gabriela Wick-Pedro, Cláudia Dias de Barros, and Oto Vale. 2024. Roda Viva boundaries: an overview of an audio-transcription corpus. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese – Vol. 2*, pages 165–169, Santiago de Compostela, Galicia, Spain. Association for Computational Linguistics.

Cláudia Freitas, Cristina Mota, Diana Santos, Hugo Gonçalo Oliveira, and Paula Carvalho. 2010. Second HAREM: Advancing the state of the art of named entity recognition in Portuguese. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Rebecca M. M. Hicke, Brian W. Haggard, Mia Ferrante, Rayhan Khanna, and David Mimno. 2025. Are You There God? Lightweight Narrative Annotation of Christian Fiction with LMs. *arXiv preprint arXiv:2507.19756*.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Anne-Stine Ruud Husevåg. 2019. From subtitles to substantial metadata: examining characteristics of named entities and their role in indexing. *International Journal on Digital Libraries*, 20(3):241–251.

Saul A. Kripke. 1972. Naming and necessity: Lectures given to the Princeton University philosophy colloquium. In *Semantics of Natural Language*, pages 253–355. Springer.

Joonatan Laato, Jenna Kanerva, John Loehr, Virpi Lummaa, and Filip Ginter. 2025. Extracting Social Connections from Finnish Karelian Refugee Interviews Using LLMs. *arXiv preprint arXiv:2502.13566*.

Jing Li, Aixin Sun, Jianglei Han, and Chenliang Li. 2020. A survey on deep learning for named entity recognition. *IEEE Transactions on Knowledge and Data Engineering*, 34(1):50–70.

Ruicheng Liu, Rui Mao, Anh Tuan Luu, and Erik Cambria. 2023. A brief survey on recent advances in coreference resolution. *Artificial Intelligence Review*, 56(12):14439–14481.

Rafael Oleques Nunes, Dennis Giovani Balreira, André Suslik Spritzer, and Carla Maria Dal Sasso Freitas. 2024. A named entity recognition approach for Portuguese legislative texts using self-learning. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese*, pages 290–300.

Thierry Poibeau. 2024. Annotating References to Mythological Entities in French Literature. *arXiv preprint arXiv:2412.18270*.

Diana Santos, Nuno Seco, Nuno Cardoso, and Rui Vilela. 2006. HAREM: An advanced NER evaluation contest for Portuguese. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA).

Elisa Terumi Rubel Schneider, João Vitor Andrioli de Souza, Julien Knafou, Lucas Emanuel Silva e Oliveira, Jenny Copara, Yohan Bonescki Gumiel, Lucas Ferro Antunes de Oliveira, Emerson Cabrera Paraiso, Douglas Teodoro, and Cláudia Maria Cabral Moro Barra. 2020. BioBERTpt - a Portuguese neural language model for clinical named entity recognition. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, pages 65–72, Online. Association for Computational Linguistics.

Fábio Capuano de Souza, Rodrigo Nogueira, and Roberto de Alencar Lotufo. 2020. BERTimbau: Pretrained BERT Models for Brazilian Portuguese. In *Intelligent Systems*, pages 403–417, Cham. Springer International Publishing.

Fábio Capuano de Souza, Rodrigo Nogueira, and Roberto de Alencar Lotufo. 2023. BERT models for Brazilian Portuguese: Pretraining, evaluation and tokenization analysis. *Applied Soft Computing*, 149:110901.

From Syntax to Semantics: Introducing UMR for NLP Annotation

Adriana S. Pagano[1], **Magali Sanches Duran**[2], **and Federica Gamba**[3]
[1] Universidade Federal de Minas Gerais, Brazil
[2] Universidade de São Paulo, Brazil
[3] Charles University, Czechia
apagano@ufmg.br magali.duran@gmail.com gamba@ufal.mff.cuni.cz

Abstract

Uniform Meaning Representation (UMR) is a cross-linguistic semantic representation framework designed to encode sentence meaning in a structured and interpretable way. Building on the foundations of Abstract Meaning Representation (AMR), UMR extends semantic coverage to events, participants, semantic roles, temporal/aspectual information, modality, and discourse links. It is language-agnostic and therefore suitable for multilingual exploration.

This tutorial provides a beginner's introduction to UMR aimed at an audience with no prior experience with AMR, UMR, or meaning representations. The tutorial begins with a simple introduction to the essentials of Universal Dependencies (UD) needed to understand how UMR graphs can be constructed from syntactic information. Using simple Portuguese examples, the tutorial illustrates how basic UD structures guide the creation of UMR graphs. Participants will leave with a foundational understanding of what UMR is; how it relates to syntax and semantic roles; how to create minimal UMR graphs, and how Portuguese UD treebanks can support UMR annotation.

Adriana S. Pagano
Universidade Federal de Minas Gerais

Adriana S. Pagano is a Full Professor of Applied Linguistics and Translation Studies at Universidade Federal de Minas Gerais (UFMG). She holds a BA in Translation (UNLP), an MA in English Language and Literature (UFSC), and a PhD in Linguistic and Literary Studies (UFMG). She has led and collaborated on several NLP projects involving translation and post-editing, natural language understanding, and natural language generation. She currently coordinates dependency syntax annotation projects in the healthcare domain and collaborates with annotation initiatives at the Center for Artificial Intelligence (C4AI). Her research interests include systemic-functional grammar, NLU/NLG, and linguistically informed approaches to annotation.

Magali Sanches Duran
Universidade de São Paulo

Magali Sanches Duran holds a degree in Translation Studies from UNESP (1985), an MBA from FGV-São Paulo (1992), and a Master's (2004) and PhD (2008) in Linguistics from UNESP. She completed multiple postdoctoral research projects at the Núcleo Interinstitucional de Linguística Computacional (NILC/USP-São Carlos), working between 2009 and 2025 on initiatives funded by Microsoft, Samsung, and IBM at the Centro de Inteligência Artificial (C4AI). Her expertise includes syntactic and semantic corpus annotation, sentiment analysis, text complexity metrics, and the development of lexical resources for NLP.

Federica Gamba
Charles University

Federica Gamba is a PhD candidate at the Institute of Formal and Applied Linguistics (UFAL), Charles University, specializing in semantic and syntactic annotation with a focus on Uniform Meaning Representation (UMR) and Universal Dependencies (UD). She has worked on multilingual resource development as a Visiting Researcher at the University of Colorado Boulder and previously as a Research Fellow at the Institute of Computational Linguistics (CNR-ILC) in Pisa. Her background includes work on lexical and textual resources in low-resource and historical languages, supported by advanced training at the University of Pavia, IUSS Pavia, and the Université Paris-Sorbonne.

Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)
Vol. 2, pages 312–312
April 13-16, 2026

Author Index