

STIL 2024

15th Symposium in Information and Human
Language Technology

Proceedings of the Conference, Vol. 1

November 18, 2024

About the workshop

The Proceedings of the XV Brazilian Symposium on Information Technology and Human Language (STIL 2024) present the selected papers presented at the event held from September 18 to 21, 2024, in the city of Belém, PA, in conjunction with the XXXIV Brazilian Conference on Intelligent Systems (BRACIS 2024), XXI National Meeting on Artificial and Computational Intelligence (ENIAC 2024), and XII Symposium on Knowledge Discovery, Mining and Learning (KDMiLe 2024). In this edition, the proceedings compile the selected papers for the Main Conference and the works accepted in its satellite events:

- XV Brazilian Symposium on Information Technology and Human Language (STIL 2024) received a total of 99 paper submissions, consisting of 78 full papers and 21 short papers. Among the full papers, 24 were accepted, resulting in an acceptance rate of 30%. As for the short papers, 13 were accepted, corresponding to an acceptance rate of 62%. These works were selected through a double-blind peer review process;
- IX Portuguese Description Conference (JDP 2024) received a total of 13 article submissions, of which 10 were accepted, resulting in an acceptance rate of 77%. These works were selected through a double-blind peer review process;
- I Workshop on Data Enrichment in Portuguese (PaDAWan 2024) received a total of 12 article submissions, of which 6 were accepted, resulting in an acceptance rate of 50%. These works were selected through a double-blind peer review process.

Acknowledgments

The Program Committee chairs acknowledge the financial support to this conference provided by the Brazilian Computer Society (SBC). We thank the Program Committees of the XIV Brazilian Symposium in Information and Human Language Technology and Collocated Events for their reviews. Last but not least, we are grateful to the local organization committee led by Carlos Renato Lisboa Francês (UFPA), Evelin Helena Silva Cardoso (UFPA), Jose Jailton Henrique Ferreira Junior (UFPA), Hugo Pereira Kuribayashi (UNIFESSPA) and Jorge Antonio Moraes de Souza (UFRA).

November 2024

Adriana Pagano (UFMG, Brazil)
Daniela Barreiro Claro (UFBA, Brazil)

Program chairs

- **STIL**

Adriana Pagano (UFMG, Brazil)
Daniela Barreiro Claro (UFBA, Brazil)

- **JDP**

Raquel Meister Ko Freitag (UFS, Brazil)
Rerisson C. de Araújo (UFBA, Brazil)

- **TILIC**

Eloize R. Marques Seno (IFSP, Brazil)
Marcio Inácio (Univer. Coimbra, Portugal)

- **PaDaWan**

Livy Real (CE-PLN/SBC)
Evandro Fonseca (Blip/PUCRS)
Paula Cardoso (UFPA)

Program Committee

Aline Paes - Universidade Federal Fluminense
Ariani Di Felippo - Universidade Federal de São Carlos
Arnaldo Candido Junior - Universidade Estadual Paulista
Carlos Ferreira - Universidade Federal de Ouro Preto
Cássio Faria da Silva - Rede Gonzaga de Ensino Superior – REGES
Christopher Shulby - Universidade de São Paulo
Clarissa Xavier - SiDi
Cláudia Dias de Barros - Inst. Federal de Edu., Ciência e Tecnologia de São Paulo
Cláudia Freitas - Universidade de São Paulo
Diana Santos - Linguateca/Universidade de Oslo
Diego Furtado Silva - Universidade de São Paulo
Eduardo Gonçalves - Escola Nacional de Ciências Estatísticas (ENCE/IBGE)
Eduardo Luz - Universidade Federal de Ouro Preto
Elisa Marchioro Stumpf - Universidade Federal do Rio Grande do Sul
Eloize Seno - Instituto Federal de Educação, Ciência e Tecnologia de São Paulo
Ely Edison Matos - Universidade Federal de Juiz de Fora
Evandro Ruiz - Universidade de São Paulo
Evelin Amorim - INESC TEC
Gabriela Wick-Pedro - Instituto Brasileiro de Informação em Ciência e Tecnologia
Helen de Andrade Abreu - Universidade Federal de Juiz de Fora
Helena Caseli - Universidade Federal de São Carlos
Heliana Mello - Universidade Federal de Minas Gerais
Hilário Tomaz de Oliveira - Instituto Federal do Espírito Santo
Jackson Souza - Universidade Federal da Bahia
Jorge Baptista - Universidade do Algarve
Juliano Antonio - Universidade Estadual de Maringá
Lívia Ruback - Universidade Estadual de Campinas
Leandro H. M. de Oliveira. - Empresa Bras. de Pesquisa Agropecuária (EMBRAPA)
Livy Real - B2W Digital/GLiC
Lucelene Lopes - Universidade de São Paulo
Magali Duran - Universidade de São Paulo
Marcelo Finger - Universidade de São Paulo
Maria das Graças V. Nunes - Universidade de São Paulo, São Carlos.

Maria José B. Finatto - Universidade Federal do Rio Grande do Sul
Marlo Souza - Universidade Federal da Bahia – UFBA
Maucha Andrade Gamonal - Universidade Federal de Minas Gerais
Márcio de Souza Dias - Universidade Federal de Catalão
Nádia Silva - Universidade Federal de Goiás
Norton Roman - Universidade de São Paulo
Oto Vale - Universidade Federal de São Carlos
Pablo Gamallo - University of Santiago de Compostela
Paula Figueira Cardoso - Universidade Federal do Pará
Rafael Anchiêta - Instituto Federal de Educação, Ciência e Tecnologia do Piauí – IFPI
Renata Vieira - Universidade de Évora
Renato Moraes Silva - Universidade de São Paulo
Ricardo Marcacini - Universidade de São Paulo
Rodrigo Wilkens - University of Exeter
Roney Santos - Universidade Federal da Bahia
Sandra Avila - Universidade Estadual de Campinas
Sergio Antonio A. Freitas - Universidade de Brasília
Thiago A. Salgueiro Pardo - Universidade de São Paulo
Tiago Timponi Torrent - Universidade Federal de Juiz de Fora
Valéria Feltrim - Universidade Estadual de Maringá
Valeria de Paiva - Topos Institute USA

Additional Reviewers

Aline Ioste - Universidade de São Paulo
Andre Coneglian - Federal University of Minas Gerais
Annie Amorim - Universidade Federal Fluminense
Arthur Scalercio - Universidade Federal Fluminense
Babacar Mane - Universidade Federal da Bahia
Bruno Cabral - Universidade Federal da Bahia
Eduardo Gonçalves - Escola Nacional de Ciências Estatísticas (ENCE/IBGE)
Elisa Marchioro Stumpf - Universidade Federal do Rio Grande do Sul
Felipe Serras - Universidade de São Paulo
Fernando Moraes - Universidade Federal da Bahia
Izabela Müller - Universidade do Algarve / INESC-ID Lisboa
Jonnathan Carvalho - Instituto Federal Fluminense – RJ
Larissa D. Xavier da Silva - Universidade Federal da Paraíba
Laura A. Costa Ribeiro - Universidade Federal Fluminense
Lilian Teixeira de Sousa - Universidade Federal da Bahia
Marcos Treviso - Universidade de São Paulo
Reginaldo Santos - Universidade Federal do Pará
Rogerio Sousa - Instituto Federal do Piauí
Sidney Leal - Universidade de São Paulo

Contents

STIL

A Linguagem em Foco: Anotação de Sinalizadores Discursivos em Textos Jornalísticos

Paula Cardoso, Jackson Souza, Roana Rodrigues, Ewerson Dantas, Larissa Santa Bárbara, Mateus Araújo, Naira Gama, Tobias Almeida, Gabriel Cruz.....1-10

Synthetic AI Data Pipeline for Domain-Specific Speech-to-Text Solutions

Anderson Luiz Karl, Guilherme Sales Fernandes, Leonardo Augusto Pires, Yvens R. Serpa, Carlos Caminha.....11-21

Automatic Annotation of Enhanced Universal Dependencies for Brazilian Portuguese

Elvis A. de Souza, Magali S. Duran, Maria das Graças V. Nunes, Gustavo Sampaio, Giovanna Belasco, Thiago A. S. Pardo.....22-31

Biases in GPT-3.5 Turbo model: a case study regarding gender and language

Fernanda Malheiros Assi, Helena de Medeiros Caseli.....32-43

Modestos e Sustentáveis: O Ajuste Eficiente Beneficia Modelos de Língua de Menor Escala em Português?

Gabriel Assis, Arthur Vasconcelos, Lívia de Azevedo, Mariza Ferro, Aline Paes.....44-54

TableRAG: A Novel Approach for Augmenting LLMs with Information from Retrieved Tables

Elvis A. de Souza, Patricia F. da Silva, Diogo Gomes, Vitor Batista, Evelyn Batista, Marco Pacheco.....55-64

PropBank e anotação de papéis semânticos para a língua portuguesa: O que há de novo?

Cláudia Freitas, Thiago Alexandre Salgueiro Pardo.....65-75

LLMs as Tools for Evaluating Textual Coherence: A Comparative Analysis

Bryan K. S. Barbosa, Claudio E. C. Campelo.....76-85

Evaluating Federated Learning with Homomorphic Encryption for Medical Named Entity Recognition Using Compact BERT Models

Marcos F. Pontes, Rodrigo C. Pedrosa, Pedro H. Lopes, Eduardo J. Luz.....86-94

A Dependency Treebank of Tweets in Brazilian Portuguese: Syntactic Annotation Issues and Approach

Ariani Di Felippo, Maria das Graças V. Nunes, Bryan K. da Silva Barbosa.....95-104

Geração Automática de Perguntas em Português do Brasil Usando os Modelos PTT5 e FLAN-T5

Tiago Felipe V. Braga, Bruno Cardoso Coutinho, Hilário Tomaz Alves de Oliveira.....105-114

Sumarização Automática de Artigos de Notícias em Português: Da Extração à Abstração com Abordagens Clássicas e Modelos de Neurais

Marcio Alves Sarmento, Hilário Tomaz Alves de Oliveira.....115-124

Boosting not so Large Language Models by using Knowledge Graphs and Reinforcement Learning

William Jones Beckhauser, Renato Fileto.....125-135

Genipapo – a Multigenre Dependency Parser for Brazilian Portuguese

Ariani Di Felippo, Norton T. Roman, Bryan K. S. Barbosa, Thiago A. S. Pardo..... 136-145

Toxic Text Classification in Portuguese: Is LLaMA 3.1 8B All You Need?

Amanda S. Oliveira, Pedro H. L. Silva, Valéria de C. Santos, Gladston Moreira, Vander L. S. Freitas, Eduardo J. S. Luz..... 146-155

Disfluency Detection and Removal in Speech Transcriptions via Large Language Models

Pedro L. S. de Lima , Cláudio E. C. Campelo..... 156-164

Detection and Censorship of Offensive Language in Extended Texts in Portuguese

Lucas Lenocho de Souza, Franciele Beal, André Roberto Ortoncelli, Marlon Marcon..... 165-174

EyetrackingMOS: Proposta de um método de avaliação online para modelos de síntese de fala

Gustavo E. Araújo, Julio C. Galdino, Rodrigo de F. Lima, Leonardo Ishida, Gustavo W. Lopes, Miguel Oliveira Jr., Arnaldo Candido Jr., Sandra M. Aluísio, Moacir A. Ponti..... 175-184

Quati: A Brazilian Portuguese Information Retrieval Dataset from Native Speakers

Mirelle Bueno , E. Seiti de Oliveira , Rodrigo Nogueira , Roberto Lotufo, Jayr Pereira..... 185-195

Mineração de Argumentos em Textos de Redes Sociais no Idioma Português

Vitor Domingos Balduino dos Santos, Livia Alabarse dos Santos, Orlando B. Coelho (in memoriam), Renata Mendes de Araujo, Ivan Carlos Alcântara de Oliveira..... 196-206

A Hybrid Machine Learning Method to Author Name Disambiguation

Natan S. Rodrigues, Célia G. Ralha.....207-216

Adapting LLMs to New Domains: A Comparative Study of Fine-Tuning and RAG strategies for Portuguese QA Tasks

Leandro Yamachita da Costa, João Baptista de Oliveira e Souza Filho.....217-227

A Change in Perspective: The Trade-Off Between Perspective API and Custom Models in Classifying Hate Speech in Portuguese

Arthur Buzelin, Yan Aquino, Pedro Bento, Samira Malaquias, Wagner Meira Jr, Gisele L. Pappa.....228-236

No Argument Left Behind: Overlapping Chunks for Faster Processing of Arbitrarily Long Legal Texts

Israel Fama, Bárbara Bueno, Alexandre Alcoforado, Thomas Palmeira Ferraz, Arnold Moya, Anna Helena Reali Costa.....237-246

Syntactic parsing: where are we going?

Lucelene Lopes, Thiago Alexandre Salgueiro Pardo, Magali S. Duran.....247-254

Segmentação Textual Baseada em Tópicos em Português Utilizando BERTimbau

Luciano A. C. da Silva, Maiara S. F. Rodrigues, Adriana P. Archanjo, Luis Pessoa, Miguel L. Silva, Thiago F. de Almeida, Leonardo Silveira,.....255-259

Avaliação de modelos para detecção de ataques de replay usando diferentes bases de dados

Giovana Y. Nakashima, Higor D. C. Santos, Jone W. M. Soares, Mário Uliani Neto, Fernando O. Runstein, Ricardo P. V. Violato, Marcus Lima.....260-265

Avaliação de arquiteturas de síntese de fala generativa com abordagens de espectrograma e fim-a-fim em cenários low-resource para clonagem de voz

Bruno C. dos S. Ribeiro, Gustavo H. dos S. Figueiredo, Leonardo H. da S. Correia, Mário Uliani Neto, Fernando O. Runstein, Ricardo P. V. Violato, Marcus Lima266-270

Leveraging Structured Data Input for Effective Chatbot Integration in Enterprises

Caio Siqueira, Orlando Guilarte, Giuliano Ferreira, Omar Leiva.....271-275

Anomaly Detection in Text Data: A Semi-Supervised Approach Applied to the Portuguese Domain

Fabio Masaracchia Maia, Anna Helena Reali Costa.....276-281

Identificação de aspectos explícitos e implícitos em críticas gastronômicas em português: avaliando o potencial dos LLMs

Luiz H. N. Silva, Eloize R. M. Seno, Rozane R. Rebechi, Helena M. Caseli Fabiano M. Rocha Júnior, Guilherme A. Faller.....282-287

Avaliação de Algoritmos de Clusterização para Agrupamento de Descrições de Produtos em Notas Fiscais Eletrônicas

Jonas Gabriel L. de Araújo, Thaís G. do Rêgo, Yuri de A. M. Barbosa.....288-293

Unified Knowledge-Graph for Brazilian Indigenous Languages: An Educational Applications Perspective

Gustavo Polleti, Fabio Cozman, Fabricio Gerardi.....294-299

A Robustness Analysis of Automated Essay Scoring Methods

Rafael T. Anchiêta, Rogério F. de Sousa, Raimundo S. Moura.....300-305

Classificação de Notícias em Português Utilizando Modelos Baseados em Transferência de Aprendizagem e Transformers

Wagner Narde, João Mendanha, Henrique Barbosa, Frederico Coelho, Bruno Santos, Luiz Torres.....306-310

Beyond Single Models: Leveraging LLM Ensembles for Human Value Detection in Text

Diego Dimer Rodrigues, Mariana Recamonde-Mendoza, Viviane P. Moreira.....311-316

JORNADA DE DESCRIÇÃO DO PORTUGUÊS

Performance in a dialectal profiling task of LLMs for varieties of Brazilian Portuguese

Raquel M. Ko Freitag, Túlio Sousa de Gois317-326

Mini-glossário do Tucumã do Pará no Município de Acará: olhares, significados e cultura da Amazônia

Eliene da S. Alves, Brayna C. dos S. Cardoso.....327-335

Desambiguação de lema e atributos morfológicos na anotação do corpus Porttinari-base

Lucelene Lopes, Magali S. Duran, Thiago Alexandre Salgueiro Pardo.....336-345

An NLP approach to impersonal –se in Brazilian Portuguese

Elvis A. de Souza, Magali S. Duran, Adriana S. Pagano346-355

Inferências baseadas em sintaxe: a anotação de sujeitos implícitos

Magali Sanches Duran, Maria das Graças Volpe Nunes, Thiago Pardo.....356-364

Notes on variation and lexical diachrony in the Parish Memories-Alentejo collection (1758)

Helena Freire Cameron, Fernanda Olival, Renata Vieira..... 365-374

A sílaba e a composicionalidade em emakhuwa (P31): análise de empréstimos do português

Francelino Wilson, Vasco Magona, Felermino Ali 375-385

Linguistic and emotional dynamics in satirical vs. real news: a psycholinguistic analysis

Gabriela Wick-Pedro, Roney Lira de Sales Santos, Oto Araújo Vale..... 386-392

Complementos de eco de adjetivos com completiva-sujeito em português do Brasil

Ryan Saldanha Martinez, Jorge Baptista, Oto Vale..... 393-402

Modelagem baseada em frames para identificação do léxico da Violência de Gênero

Lorena Tasca Larré, Tiago Timponi Torrent..... 403-412

WORKSHOP DE IC EM TECNOLOGIA DA INFORMAÇÃO E DA LINGUAGEM HUMANA

Relações Retóricas de List e Sequence em textos jornalísticos

Tobias J. A. Almeida, Patrícia V. Almeida, Paula C.F. Cardoso 413-417

Estudo preliminar sobre sinalizadores discursivos para Conteúdo Gerado por Usuários

Naira Silva Gama, Jackson Wilke da Cruz Souza 418-423

Relações de coerência do português brasileiro: um estudo bibliográfico-documental da RST e seus sinalizadores discursivos

Larissa Jesus Santa Bárbara, Roana Rodrigues, Jackson Wilke da Cruz Souza.....424-428

Subsídios Linguísticos para classificação automática de textos de User-Generated Content

Mateus Araújo Pereira, Jackson Wilke da Cruz Souza.....429-433

Relações de coerência do espanhol peninsular: Um estudo bibliográfico-documental da Rhetorical Structure Theory

Ewerson Dantas, Roana Rodrigues Jackson Wilke da Cruz Souza.....434-439

Estratégias automáticas para análise da concordância da anotação de Sinalizadores Discursivos

Gabriel Sizinio Bomfim Cruz, Jackson W. C. Souza, Paula C. F. Cardoso.....440-444

Mineração de Emoções Multirrótulo Em Textos Curtos

Ramon N. Mendes, Syanne. K. M. Tavares, Luiz Nicollas M. Campos, Fabíola P. O. Araújo.....445-450

Classificação automática de textos de User-Generated Content utilizando Aprendizagem de Máquina Supervisionado

Iolanda Victoria Morais Ramos, Jackson Wilke da Cruz Souza.....451-456

PLN e Segurança Jurídica Identificação de divergências jurisprudenciais com Processamento de Linguagem Natural

Marcella Queiroz de Castro; Ana Régia Mendonça.....457-462

Um Pipeline de Pré-Processamento de Dados Textuais em Português para Análise de Redes Sociais

Livia A. dos Santos, Orlando B. Coelho (in memoriam)¹, Renata Araujo, Ivan Carlos A. Oliveira.....463-468

Especulação Mística. Uma abordagem de Clusterização e Busca Semântica na aproximação de preço em cartas de Magic: The Gathering

Rodrigo Marques Duarte, André de Lima Salgado, Paula Figueira Cardoso.....469-473

Comparação de Ferramentas para Análise de Sentimentos Aplicada no Contexto Educacional

Benjamin G. Moreira, Luiz C. Camargo, Ricardo J. Pfitscher, Tatiana R. Garcia.....474-478

Modelo de Linguagem Quantizados na Área da Saúde: Um Enfoque em Perguntas e Respostas com Base na Técnica DPO

Mário Pinto Freitas Filho, João Dallyson Sousa de Almeida, Anselmo C. Paiva.....479-483

PORTUGUESE DATA AUGMENTATION WORKSHOP

LLM-SEMREL: Towards a Better Coreference Resolution for Portuguese

Evandro Fonseca, Joaquim Neto..... 484-492

Automated Topic Annotation in Brazilian Product Reviews: A Case Study of Adversarial Examples with Sabia-3

Lucas Nildaimon dos Santos Silva, Livy Real..... 493-501

Text extraction from Knowledge Graphs in the Oil and Gas Industry

Laura P. Navarro, Elvis A. de Souza, Marco A. C. Pacheco 502-507

Getting Logic From LLMs Annotating Natural Language Inference with Sabiá

Fabiana Avais, Marcos Carreira, Livy Real 508-517

Augmenting Data to Improve the Performance of Recommender Systems

Leticia Freire de Figueiredo, Joel Pinho Lucas, Aline Paes 518-521

Brazilian Consumer Protection Code: a methodology for a dataset to Question-Answer (QA) Models

Aline Athaydes, Lucas Bulcao, Caio Sacramento, Babacar Mane, Daniela Barreiro Claro, Marlo Souza, Robespierre Pita..... 522-529

A Linguagem em Foco: Anotação de Sinalizadores Discursivos em Textos Jornalísticos

Paula Cardoso¹, Jackson Souza², Roana Rodrigues³, Ewerson Dantas³,
Larissa Santa Bárbara³, Mateus Araújo², Naira Gama²,
Tobias Almeida⁴, Gabriel Cruz²

¹Universidade Federal do Pará - Belém/PA

²Universidade Federal da Bahia - Salvador/BA

³Universidade Federal de Sergipe - Aracajú/SE

⁴Universidade Federal de Lavras - Lavras/MG

pcardoso@ufpa.br, {jacksoncruz, pereiramateus, gabrielsizinio, nairagama}@ufba.br,
{roana, larissajesus}@academico.ufs.br, tobias.almeida@estudante.ufla.br

Abstract. *Due to their capability to enable the observation of linguistic and social behaviors, annotated corpora have become of interest to various fields of study. In the context of Rhetorical Structure Theory (RST), this paper presents the methodological and practical processes involved in annotating discourse markers within a Brazilian Portuguese journalistic corpus. Additionally, we provide initial quantitative and qualitative assessments of the decisions made by the annotation team.*

Resumo. *Por serem recursos que permitem a observação de comportamentos e usos linguísticos e sociais, os corpora anotados passaram a ser de interesse de diferentes áreas do conhecimento. No contexto da Rhetorical Structure Theory (RST) apresentamos neste trabalho os processos metodológicos e práticos de anotação de sinalizadores discursivos em um corpus jornalístico do português do Brasil. Ainda, apresentamos as primeiras avaliações (quanti e qualitativa) sobre as decisões tomadas pelo grupo de anotadores.*

1. Introdução

A Linguística de *corpus* (LC), enquanto área, instiga a utilização de técnicas e metodologias que nos levam a reunir grandes conjuntos de dados textuais (escritos, orais ou multimodais), a fim de descrever fenômenos linguísticos. Em interface ao Processamento de Linguagem Natural (PLN), uma das tarefas que a LC se propõe a realizar é a anotação desse conjunto de dados, tida como “o processo de enriquecer um *corpus*, adicionando informações linguísticas inseridas por humanos ou máquinas com um objetivo teórico ou prático” [Pedro e Vale 2018].

Por serem recursos que permitem a observação de comportamentos e usos linguísticos e sociais, os *corpora* anotados passaram a ser de interesse de diversas áreas do conhecimento, como Humanidades digitais, Linguística e Computação.

[Pustejovsky e Stubbs 2012] apontam que a análise dos *corpora* permite desvendar a natureza da linguagem e, conseqüentemente, capturar possíveis propriedades que possam ser modeladas computacionalmente.

Porém, esse processo de anotação tende a ser custoso, já que grande parte é realizada de forma semiautomática e requer intervenção humana. [Hovy e Lavid 2010] apresentam uma metodologia genérica sobre esse processo, que engloba etapas como preparação do conjunto de dados, instanciação da base teórica, anotação de fragmentos do *corpus*, medição das decisões de anotação e escalabilidade do processo de maneira automática. No entanto, essa tarefa pode ser ajustada conforme o tipo de anotação a ser realizada, o que pode levar à omissão de algumas das etapas sugeridas pelos autores.

[Taboada e Das 2013] e [Liu e Zeldes 2019], a partir de *corpora* pré-annotados com o modelo *Rhetorical Structure Theory* (RST) [Mann e Thompson 1987] identificaram uma série de pistas linguísticas e estruturais que serviam de sinalizadores para as relações discursivas previamente identificadas. Ambos os trabalhos organizaram os Sinalizadores Discursivos (SD) em função de suas características (semânticas ou sintáticas, por exemplo), pondo em xeque a ideia de que as relações RST deveriam ser identificadas majoritariamente por meio de Marcadores Discursivos (MD), tidos como preposições e conjunções.

Com base nessa metodologia, [Rodrigues et al. 2023] descreveram SDs para além dos MDs a partir do *corpus* CSTNews [Cardoso et al. 2011]. Tal recurso linguístico-computacional consiste em um conjunto de textos jornalísticos em Português que já havia sido anotado segundo o modelo RST. A RST propõe que um texto coerente é formado por unidades mínimas de discurso (*Elementary Discourse Units* - EDU ou proposições) que desempenham funções retóricas para que o objetivo comunicacional do autor seja atingido. Partindo dessa anotação prévia, os anotadores deste trabalho, por sua vez, identificaram apenas os sinalizadores que consideraram relevantes para caracterizar e/ou indicar determinadas relações, como em (1) - extraído do *corpus* CSTNews.

(1) [A seleção brasileira masculina de vôlei,]^A [que é treinada por Bernardinho,]^B [venceu a Finlândia por 3 sets a 0.]^C

As porções (1a) e (1c) foram conectadas por meio da relação RST *Same-Unit*, indicando que se trata da mesma unidade, já que estão separadas por haver detalhamento informacional em 1B em relação à (1a) por meio da relação *Elaboration*. [Rodrigues et al. 2023] indicaram que a pontuação (no caso, vírgula), a concordância verbal e o encaixamento de outra relação RST poderiam ser utilizadas como pistas para a identificação da relação *Same-unit*.

Esse estudo preliminar resultou em um manual de anotação de SDs em textos jornalísticos [Dantas et al. 2024], em que há, além de instruções, a proposta da primeira taxonomia de sinalizadores de relações RST para o PB. Destaca-se que esse tipo de recurso com explicações, exemplos e instruções objetivas subsidia a decisão dos anotadores diante de fatos novos e/ou já conhecidos [Duran et al. 2022].

Assim, objetivamos neste trabalho relatar as etapas metodológicas e práticas de anotação de SDs no *corpus* CSTNews, além de apontar as primeiras avaliações sobre as decisões tomadas pelo grupo de anotadores. Para tanto, este trabalho está organizado em 5 seções, além desta Introdução. Na Seção 2, destacamos trabalhos relacionados ao

processo de anotação e análise em RST, sobretudo para o PB. Na Seção 3, detalhamos a metodologia de anotação empreendida neste estudo. Na Seção 4 apresentamos os resultados e as discussões correspondentes. Por fim, na Seção 5, tecemos algumas considerações finais.

2. Trabalhos Relacionados

Identificar relações RST por meio de marcas explícitas no texto não é uma tarefa nova, especialmente em PLN para análise de discurso. Os MDs são tidos como conectivos entre porções textuais, sinalizando determinadas relações discursivas, como o “mas” para oposição, por exemplo. A análise das relações discursivas (ou de coerência) está intimamente ligada a descobrir a intenção do autor ao apresentar partes do texto em uma ordem e combinação específicas. Portanto, trata-se de uma tarefa que vai além de identificar os MDs.

A literatura [Marcu 2000, Pardo 2005, Taboada e Das 2013] indica que a identificação de MD, em função da relação RST a que ocorrem, facilita o processamento do texto. Estudos recentes [Das e Taboada 2018, Liu e Zeldes 2019] afirmam que os MDs sinalizam apenas um número restrito de relações dentro de um texto, e sugerem que as relações RST podem ser identificadas por sinais que vão além deles. Como os MDs não marcam explicitamente as relações e não são exclusivos, a noção de SD parece ser mais apropriada do que a de MD nesse contexto.

[Das e Taboada 2018] argumentam que para uma comunicação ser eficaz, é fundamental que as relações sejam interpretadas de maneira relativamente clara, o que requer sinalizadores precisos. Os autores acreditam que as relações de coerência são entidades cognitivas, e, portanto, é possível descobrir como ouvintes e leitores as identificam usando indicadores que auxiliem o processo interpretativo. Utilizando o *RST Discourse Treebank*, os autores realizaram uma anotação detalhada dos SD, resultando no *RST Signalling Corpus* (RST-SC). Eles observaram que pode haver relações sinalizadas por um único sinalizador (como MD, referências pessoais, orações relativas ou dois pontos) ou por combinações de SD (como vírgula + oração no particípio passado, ou construção sintática paralela + cadeia lexical). Quando surgia uma nova instância de um tipo específico de relação, os anotadores consultavam a taxonomia para encontrar o(s) sinalizador(es) mais adequado para aquela instância. Durante o processo de anotação, os autores observaram casos em que não foi possível determinar com precisão o SD que representava uma determinada relação.

Em [Liu e Zeldes 2019] descreve-se um esforço de anotação para ancorar SD a partir de diversas categorias tais como sintática, semântica, gráfica and morfológica. Seus resultados mostraram que, com 11 documentos e 4.732 *tokens*, 923 foram instâncias de SD, o que representou mais de 92% dos sinais discursivos. O tipo semântico representou a maioria dos casos, enquanto as relações discursivas ancoradas por DM corresponderam a apenas cerca de 8,5% dos tokens ancorados.

Quanto à língua portuguesa, [Pardo 2005] foi o precursor em investigar a construção de analisadores discursivos. A partir de um *corpus* de textos científicos e anotado com RST, o autor identificou diversos padrões de análise que especificam os relacionamentos entre as relações retóricas e seus marcadores textuais. Apesar de muitos padrões serem baseados em MD, o autor ressalta que não existe uma relação *sine qua non*

entre MD e as relações que sinalizam, pois uma mesma relação pode ser sinalizada por vários marcadores (por exemplo, a relação *Concession* pode ser sinalizada pelos marcadores “entretanto”, “no entanto”, entre outros) e um mesmo marcador pode sinalizar várias outras relações (por exemplo, o marcador “porque” pode sinalizar as relações *Cause* e *Result* (volitivas ou não), *Justify*, *Explanation*, entre outras).

Ainda com relação ao português, [Maziero 2016] investigou atributos de organização textual, da morfossintaxe, da sintaxe, da semântica e discurso para construir um analisador discursivo baseado na RST. A partir da análise de *corpora* anotados com RST, o autor aponta que: a) existem relações que apresentam grande subjetividade, tais como as relações *Evidence*, *Justify* e *Explanation*; b) a relação *Same-unit* ocorre apenas no nível intrassentencial, o que é esperado, pois é responsável por ligar proposições quebradas por uma relação de *Parenthetical* ou *Elaboration*, por exemplo; c) algumas relações são mais frequentes no nível intrassentencial do que no inter-sentencial. Após vários experimentos com aprendizado de máquina para identificação das relações discursivas no nível intrassentencial, o autor concluiu que atributos morfossintáticos proporcionaram melhores resultados do que os atributos semânticos e discursivos.

3. Metodologia

A anotação de SDs foi feita a partir do *corpus* CSTNews¹. Os textos do *corpus* estão organizados em 50 conjuntos, com dois ou três documentos que noticiam o mesmo evento. Por essa característica multidocumento e de redundância, a anotação foi feita apenas no maior texto do conjunto, pois acreditamos que quanto maior for o texto, maior é a chance de encontrarmos mais relações RST e, possivelmente, essas relações ocorram nos outros textos da mesma coleção. Nesse caso, foram separados 50 textos para esta tarefa de anotação.

A anotação de SDs foi realizada por meio da ferramenta *rstWeb* [Zeldes 2016], que é uma plataforma desenvolvida para facilitar a análise e a anotação de textos com base na RST. Essa ferramenta permite aos usuários realizar análises estruturais detalhadas dos textos, identificando proposições e suas relações de coerência conforme proposto pela teoria. Neste trabalho, a taxonomia de SDs, na Figura 1, foi implementada.

Os textos escolhidos foram pré-processados e distribuídos a um grupo de oito pesquisadores. A anotação aconteceu de maneira assíncrona, em que cada anotador recebia semanalmente 3 ou 4 textos. Cada texto foi anotado por três anotadores para que pudessemos ter uma versão do *corpus* com a decisão sobre a indicação dos SDs por maioria simples.

Para promover discussão e resolução de dúvidas, especialmente sobre casos não previstos pelo manual, foram conduzidas reuniões semanais com o grupo. Além disso, dois dos anotadores, por terem mais experiência com tarefas nesse sentido, nunca ficavam juntos no trio, para que pudessem auxiliar na resolução de dúvidas de maneira assíncrona. Ressalta-se que os anotadores possuíam diferentes formações acadêmicas (linguistas ou cientistas da computação) e com experiências distintas em tarefas de anotação de *corpus*. Por conta disso, foi necessária uma etapa de treinamento para que o grupo se familiarizasse com a taxonomia de SDs e com o modelo RST, além de ter acesso ao manual de

¹Disponível em: <http://nilc.icmc.usp.br/CSTNews/login/?next=/CSTNews/>

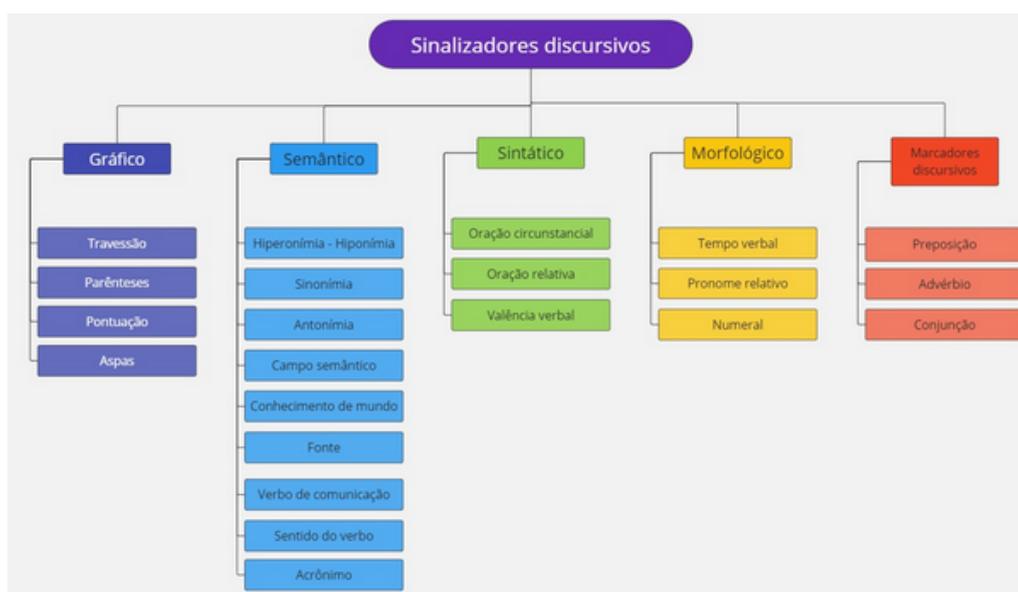


Figura 1. Taxonomia de sinalizadores discursivos proposta por Autores (2023).

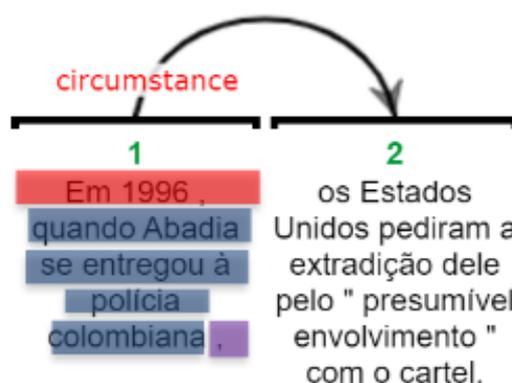


Figura 2. Taxonomia de sinalizadores discursivos proposta por Autores (2023).

anotação para auxiliar em suas decisões.

Na fase de treinamento, foi realizada a anotação de três textos do *corpus*. Em reuniões síncronas, os anotadores puderam corrigir possíveis equívocos e identificar quais unidades do discurso deveriam ser devidamente anotadas, ou seja EDUs que estivessem presentes em um mesmo período sintático.

Na Figura 2, têm-se um exemplo de anotação para a relação *Circumstance*. Essa relação RST deve apresentar uma situação realizável, em que o satélite (EDU 1, provê a situação que é apresentada no núcleo - EDU 2). No exemplo, tem-se que essa relação está sendo sinalizada por meio de Advérbio (vermelho), Oração circunstancial (azul) e Pontuação (lilás).

Três textos de diferentes tamanhos foram anotados por todos os anotadores em distintas fases do processo, com o objetivo de medir periodicamente a concordância do grupo. Esse processo se repetiu a cada 10 conjuntos de textos anotados. Ao final de 2 meses, 47 textos foram anotados.

Trecho anotado com a relação RST Concession	Anotadores	Sinalizadores indicados
“Tal capacidade de mutação fez escola, mas dificilmente as criaturas saberão superar o criador.”	A	“mas” + “,”
	B	“mas” + “,”
	C	“mas”

Tabela 1. Comparação de anotações.

A concordância foi medida automaticamente a partir de duas abordagens. Na abordagem *gold* observa-se estritamente o que o grupo de anotadores apontaram como sinalizador, sendo, portanto, mais restrita. Já na abordagem *silver* definiu-se um intervalo de cinco janelas (à esquerda e à direita) em relação ao sinalizador-alvo, como demonstrado na Tabela 1.

Na Tabela 1, tem-se um exemplo de como as duas abordagens da concordância foram aplicadas. Os anotadores A e B indicaram os mesmos sinalizadores, ao passo que o sinalizador C indicou apenas um em comum com o grupo. Caso fosse considerada apenas uma análise mais restritiva sobre a concordância, a decisão do anotador C prejudicaria o cálculo, ao passo que numa abordagem mais ampla, sua decisão não traria tantos prejuízos.

Apesar de todos os esforços metodológicos de distribuição de textos, e de todos os anotadores estarem alinhados junto ao modelo teórico e à ferramenta utilizados, é possível que fatores externos à tarefa influenciam na disposição dos anotadores, fazendo-os eventualmente não apenas discordarem sobre um sinalizador, mas também não se atentarem a realizar a indicação adequada. Por conta disso, escolheu-se neste trabalho não apenas realizar uma análise mais restrita sobre a concordância, mas também mais ampla, admitindo-se nesta, sobretudo, a dimensão mais subjetiva da tarefa.

Em ambas as abordagens utiliza-se a medida *Krippendorff Alpha* [Krippendorff 2011]. Trata-se de uma medida que avalia a concordância entre dois ou mais anotadores, o que se encaixa melhor no contexto deste trabalho, já que cada texto foi anotado por três pessoas. O resultado da concordância é medido num intervalo que varia entre -1 e 1, em que valores mais próximos a 1 indicam alta concordância; valores próximos a 0 indicam baixa concordância; e valores próximos a -1 indicam discordância total.

4. Resultados e Discussão

Na Tabela 2, tem-se a média dos resultados das concordâncias *gold* e *silver* da anotação em diferentes etapas do processo. Como dito, na fase de treinamento (*clusters* 1, 2 e 3), os anotadores realizaram uma primeira anotação e, após reunião de alinhamento, fizeram correções. O cálculo da concordância geral (*clusters* 16, 31 e 39) foi feito sobre o mesmo texto anotado por todo o grupo.

Dado que o processo de anotação pode ser longo e complexo, fatores externos aos aspectos linguísticos (como cansaço e diminuição da atenção, por exemplo) pode ter influenciado os anotadores. É possível perceber isso ao comparar as fases de treinamento com as demais, em que as demais sofreram decréscimos discretos. Além disso, outro possível aspecto que pode ter influenciado nesse resultado é a distribuição das relações

Fase do trabalho	Concordância	
	Gold	Silver
Treinamento	0,581	0,693
Concordância geral	0,460	0,596
Rodadas de anotação	0,496	0,691

Tabela 2. Resultado da concordância.

RST no *corpus* CSTNews. [Cardoso et al. 2011] apontam que há relações RST que ocorrem apenas uma vez, como *Otherwise*, por exemplo, e outras que aconteceram de maneira predominante, como *Elaboration*, que ocorreu 1,514 vezes. Nesse caso, é possível que, ao se deparar com uma relação RST não prevista na fase de treinamento e, portanto, ausente no manual de anotação, os anotadores enfrentaram dificuldades em indicar possíveis sinalizadores das relações em questão.

Além de uma análise quantitativa, foram feitas observações qualitativas preliminares. Para tanto, durante a anotação, os anotadores realizaram indicações de dúvidas, inconsistências e/ou outras questões em um formulário eletrônico. Ao final de cada semana, todos os apontamentos eram compilados e discutidos entre o grupo para aprimorar o processo. A partir disso, é possível destacar alguns pontos:

a) Considerações sobre o processo de anotação

Em caso de não encontrar uma etiqueta para representar o fenômeno observado, o anotador poderia registrar os *tokens* envolvidos e marcar como CPD (Casos Para Discutir depois). Em discussões e análises preliminares, os anotadores destacaram a intenção de marcar as entidades mencionadas no segmento textual. [Das e Taboada 2018], por sua vez, descrevem que os anotadores discordavam bastante entre entidade e tipos semânticos, ou seja, enquanto um anotador seleciona entidade como o sinal relevante para uma certa relação, o outro anotador a anota como sendo semântica. Os autores observaram que muitos dos atributos de entidade e características semânticas na verdade se sobrepõem. Dessa forma, essa dificuldade acontece também para a língua inglesa.

Assim como [Liu 2019, Das e Taboada 2018] relatam, também observamos no *corpus* de estudo várias relações que não tinham um token explícito para servir de sinalizador. Esses casos foram registrados como CPD. Por outro lado, as primeiras análises revelaram que alguns SD são altamente indicativos, enquanto outros são genéricos ou ambíguos. Assim, para obter uma compreensão mais precisa, é necessário considerar os contextos ao redor dos SD para desambiguá-los.

b) Considerações sobre dificuldades e limitações encontradas

A anotação das relações RST é um processo que se baseia na interpretação do analista. Assim, a depender dessa interpretação serão indicadas determinadas relações RST em detrimento de outras, resultando, então, em diferentes sinalizadores para essas relações. Neste estudo, a identificação de SDs foi feita por um grupo majoritariamente diferente de quem fez a anotação RST, com uma distância temporal considerável entre as duas tarefas. Esse fato, portanto, pode ter sido um dificultador para o grupo que fez a indicação dos sinalizadores.

Além disso, os anotadores destacaram que algumas relações RST utilizadas no

CSTNews são mais difíceis de interpretar, e consequentemente, torna-se um desafio apontar SD específicos, como exemplificado em (2).

(2) (...) [com o Programa Fome Zero, conseguiu atingir o primeiro ponto das Metas do Milênio - erradicar a fome -, com dez anos de antecedência,]^A [reduzindo em mais da metade a pobreza extrema.]^B

O trecho (2b) em relação ao trecho (2a) apresenta a relação *Volitional result*, ou seja, o resultado ocasionado foi não intencional. Nesse caso em específico, os anotadores indicaram que o sentido do verbo “reduzindo” seria o indicativo do resultado, porém sem menção ao aspecto volitivo. Destaca-se que a maioria dos *rols* de relações para outras línguas não prevêem diferença nesse aspecto.

Outro aspecto que parece ter apresentado dificuldade aos anotadores foi o fato de o manual de anotação ter sido desenvolvido com base no estudo de [Rodrigues et al. 2023] e os resultados da fase de treinamento. Como citado, relações e sinalizadores que não estavam previstos e que ocorreram ao longo do *corpus* podem ter ocasionado certos equívocos entre os anotadores. Ademais, o fato de o manual indicar certa correlação entre SDs e relações pode ter condicionado o olhar dos anotadores, como demonstrado em (3).

(3) [nesta terça deve se encontrar com o relator do caso na Câmara, deputado José Carlos Araújo (PR-BA)]^A [para tratar do assunto.]^B

De acordo com [Cardoso et al. 2011], a sentença entre (3a) e (3b) é de *Purpose*. O manual de anotação de SDs utilizou esse exemplo e indicou que a preposição “para” pode ser utilizada para identificar essa relação. Entretanto, o objetivo entre os segmentos pode também ser evidenciado por meio de “oração final” presente em (3b). Nesse caso, é possível que os anotadores tenham sido condicionados a partir de determinados pressupostos sobre as relações, ainda que tenham sido estimulados a indicarem em formulário eletrônico outros possíveis SDs e definições não previstos no manual.

Por fim, cabe pontuar que no repositório *online* do projeto de pesquisa “RST além dos marcadores discursivos”² disponibilizamos para consulta o *corpus* com a versão unificada entre os anotadores, a anotação de SDs e a planilha completa da concordância dos anotadores.

5. Considerações Finais

Neste trabalho buscamos detalhar a metodologia empregada na identificação de SDs em textos jornalísticos a partir da taxonomia proposta por [Dantas et al. 2024]. Destacamos que um estudo com essa abordagem em PB ainda não havia sido realizado, ao contrário do que já ocorre em outros idiomas, especialmente o inglês.

Os resultados relatados podem subsidiar outras análises em estudos futuros. Um desses estudos se concentra na investigação quali e quantitativa da correlação entre SDs e as relações RST, algo já iniciado por [Rodrigues et al. 2023] e tal como outros trabalhos fizeram [Liu 2019, Das e Taboada 2018, Pardo 2005]. Outro estudo será em relação à concordância de aspectos da anotação, como tipos (sintático e semântico, por exemplo) e subtipos (pronomes relativos e conhecimento de mundo, por exemplo) dos sinalizadores. Ao final desses estudos será possível fazer o levantamento da distribuição dos SDs no

²Disponível em <https://sites.google.com/view/rst-poetisa/>

corpus, bem como observar quais são mais ou menos consensuais entre os anotadores.

Dados os apontamentos críticos realizados sobre as limitações identificadas, destaca-se que este trabalho apresenta potencial de servir de diretriz de investigações de análises sobre as relações RST e seus SDs e aprimoramento de ferramentas e recursos para anotação de *corpus*. Tais aspectos são de extrema importância ao alargar a anotação a escalas maiores buscando não apenas ampliar a quantidade de textos, mas também diversificar os gêneros textuais a serem considerados.

6. Agradecimentos

Este trabalho foi realizado no âmbito do Centro de Inteligência Artificial da Universidade de São Paulo (C4AI - <http://c4ai.inova.usp.br/>), com o apoio da Fundação de Amparo à Pesquisa do Estado de São Paulo (processo FAPESP #2019/07665-4) e da IBM. Este projeto também foi apoiado pelo Ministério da Ciência, Tecnologia e Inovações, com recursos 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

- Cardoso, P. C., Maziero, E. G., Jorge, M. L. C., Seno, E. M., Di Felippo, A., Rino, L. H. M., Nunes, M. d. G. V., e Pardo, T. A. (2011). CSTNews - A discourse-annotated corpus for single and multi-document summarization of news texts in Brazilian Portuguese. In *Proceedings of the 3rd RST Brazilian Meeting*, pages 88–105.
- Dantas, E., Bárbara, L. d. J. S., Pereira, M. A., Gama, N. S., Almeida, T. J. A., Souza, J. W. d. C., Cardoso, P. C. F., e Rodrigues, R. (2024). *Manual de anotação de sinalizadores discursivos em textos jornalísticos*. Série de Relatórios Técnicos do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo.
- Das, D. e Taboada, M. (2018). RST signalling corpus: A corpus of signals of coherence relations. *Language Resources and Evaluation*, 52:149–184.
- Duran, M. S., Nunes, M. d. G. V., Lopes, L., e Pardo, T. A. S. (2022). Manual de anotação como recurso de processamento de linguagem natural: o modelo universal dependencies em língua portuguesa. *Domínios de Linguagem*, 16(4):1608–1643.
- Hovy, E. e Lavid, J. (2010). Towards a ‘science’ of corpus annotation: a new methodological challenge for corpus linguistics. *International Journal of Translation*, 22(1):13–36.
- Krippendorff, K. (2011). Computing krippendorff’s alpha-reliability.
- Liu, Y. (2019). Beyond the Wall Street Journal: Anchoring and comparing discourse signals across genres. *arXiv preprint arXiv:1909.00516*.
- Liu, Y. e Zeldes, A. (2019). Discourse relations and signaling information: Anchoring discourse signals in RST-DT. *Society for Computation in Linguistics*, 2(1).
- Mann, W. C. e Thompson, S. A. (1987). *Rhetorical Structure Theory: A theory of text organization*. University of Southern California, Information Sciences Institute Los Angeles.

- Marcu, D. (2000). The rhetorical parsing of unrestricted texts: A surface-based approach. *Computational linguistics*, 26(3):395–448.
- Maziero, E. G. (2016). *Análise retórica com base em grande quantidade de dados*. PhD thesis, Universidade de São Paulo.
- Pardo, T. A. S. (2005). *Métodos para análise discursiva automática*. PhD thesis, Universidade de São Paulo.
- Pedro, W. e Vale, O. (2018). Comentcorpus: o uso de mecanismos linguísticos na detecção de ironia e sarcasmo para o português do Brasil em um corpus opinativo. *Linguística de corpus: perspectivas*. Porto Alegre: Instituto de Letras da Universidade Federal do Rio Grande do Sul, pages 19–40.
- Pustejovsky, J. e Stubbs, A. (2012). *Natural Language Annotation for Machine Learning: A guide to corpus-building for applications*. O’Reilly Media, Inc.
- Rodrigues, R., Souza, J. W., e Cardoso, P. C. F. (2023). Sinalizadores retórico-discursivos: revisitando a anotação RST no cópús CSTnews. In *Anais do XIV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 249–257. SBC.
- Taboada, M. e Das, D. (2013). Annotation upon annotation: Adding signalling information to a corpus of discourse relations. *Dialogue & Discourse*, 4(2):249–281.
- Zeldes, A. (2016). rstWeb-a browser-based annotation interface for Rhetorical Structure Theory and discourse relations. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 1–5.

Synthetic AI Data Pipeline for Domain-Specific Speech-to-Text Solutions

Anderson Luiz Karl¹, Guilherme Sales Fernandes¹, Leonardo Augusto Pires¹,
Yvens R. Serpa^{1,3}, Carlos Caminha²

¹Audo Tecnologia e Saúde

²Universidade Federal do Ceará, UFC, Brasil

³Saxion University of Applied Sciences, Enschede, Netherlands

Abstract. *In this article, we propose a pipeline to fine-tune domain-specific Speech-to-Text (STT) models using synthetic data generated by Artificial Intelligence (AI). Our methodology eliminates the need for manually labelled audio data, which is expensive and difficult to obtain, by generating domain-specific data with a Large Language Model (LLM) combined with multiple Text-to-Speech (TTS) solutions. We applied our pipeline to the radiology domain and compared the results with different approaches based on the availability of domain-specific data, varying from the total absence of domain-specific data to the use of only domain-specific high-quality data (ground truth). Our performance improved the accuracy of the baseline by 40.19% and 10.63% for the WhisperX Tiny and Small models, respectively, which, although performed worse than the results from using the ground truth, shows that it is possible to achieve good results with minimal cost and effort. Finally, the result analysis shows a good insight into the amount of action necessary to achieve good results based on the availability of real data.*

1. Introduction

Automatic audio transcription, commonly referred to as Speech-to-Text (STT), has been a common practice for many work fields, such as health, justice, education, and business [Kumar 2024]. However, precision in recognizing and transcribing language is important to guarantee the correct and efficient use of the transcribed information. That is especially important in domain-specific applications, in which the use of technical terms and jargon increases the recognition and transcription challenge [Suh et al. 2024]. However, many of the typically available solutions for this problem are built on generic data. Due to that, their results are of lower quality when used in domain-specific scenarios [Chan et al. 2016].

A common approach to solving this issue is to build and refine solutions using domain-related contexts, vocabularies and other types of data [Huang et al. 2020]. Nowadays, it is standard to use generic AI models as the base for STT solutions and fine-tune these models with domain-specific data [Mak et al. 2024]. However, the fine-tuning process is expensive and requires a significant amount of data and effort [Hu et al. 2022]. For medical applications, for example, it is necessary to collect sensitive data, have health professionals check, correct and validate it, and guarantee its privacy and security in regard to the involved patients and personnel [Johnson et al. 2014].

Nevertheless, the need for high-quality STT solutions is evident in many work sectors. In Radiology, for example, it is a common practice to have physicians use STT

tools in their work practice to increase productivity over traditional transcription, the latter in which the professional records a report via voice to be later transcribed manually by another professional (usually without a medical background) [Hammana et al. 2015]. Any errors or delays in this process may result in possible harm and consequences to the patients and their treatments [Vorbeck et al. 2000]. Another common example is courts and judicial procedures, in which a large quantity of domain-specific texts is generated and often transcribed manually, resulting in expensive and inefficient processes [da Cruz et al. 2022].

In this context, this work proposes a low-cost pipeline for the training and fine-tuning of STT AI models when domain-specific data is required but not readily available. Our pipeline is based on the use of AI models to generate synthetic domain-specific data. For that, we have used a Large Language Model (LLM) to produce domain-specific content that simulates real use cases. Specifically for this work, we have explored the radiology domain, generating data for synthetic radiology reports using an LLM and a specific prompting approach. The synthetic data is then converted into audio files through Text-to-Speech (TTS) tools. Thus, the fine-tuning process is done entirely using synthetic data generated via AI. Additionally, due to the focus on being a low-cost solution, the results of this work were done by using inexpensive or freely available solutions. Simultaneously, this work also presents a comparison analysis of a range of possible final results depending on the availability of domain-specific data.

2. Related Work

Automatic audio transcription has been a fruitful research field in computer sciences over many years [Yu et al. 2010, Blackley et al. 2019]. Many of the traditional works in this field are focused on the inherited challenges of it, such as handling language subtleties, structure, and fluency [Gontier et al. 2021], and the limitations on the access of adequate datasets [Hu et al. 2022]. These challenges increase when dealing with domain-specific scenarios [Samarakoon et al. 2018].

In regards to datasets, the majority of works in the field use datasets in the English language [Casanova et al. 2022]. When working in scenarios with other languages, researchers must not only solve the recurrent STT challenges but also adapt their solutions, such as done by Gruzitis *et al.* [Gruzitis et al. 2022] which adapted their models to the Latvian language, and the work of Vivancos-Vincente *et al.* [Vivancos-Vicente et al. 2016] for Spanish and Portuguese. Alternatively, the work proposed by Casanova *et al.* [Casanova et al. 2022] shows an alternative to training models for different languages based on data augmentation from only one speaker for the targeted language, using cross-lingual voice conversion and multi-speaker TTS techniques.

Moreover, access to good domain-specific datasets is a challenge, and its production involves high costs with domain experts, data analysis, and validation. This problem is often faced with the use of synthetic data [Li et al. 2018, Rosenberg et al. 2019, Laptev et al. 2020, Huang et al. 2020, Yang et al. 2023]. However, synthetic data is frequently distant from real use cases due to the absence of mistakes and imperfections that are often common in human-made data, which makes it “too perfect” compared to real-world cases. This “perfection problem” is handled with the introduction of synthetic errors and imperfections, such as done by the Synt++ solution proposed by Hu *et al.*

[Hu et al. 2022], in which noise and random artefacts are introduced to the synthetic data generation so it more closely resembles real-life data.

Only recently the process of data synthesis using LLM have been explored, such as the work presented by Vásquez-Correa *et al.* [Vásquez-Correa et al. 2023], which generates domain-specific synthetic data through prompting to fine-tune an STT solution for the English, Spanish, and Basque languages. Silva *et al.* [Silva et al. 2024] also uses an LLM to generate synthetic data for a hardware failure prediction dataset. Their dataset was generated from problem categories and reports from major component manufacturers in the market.

Similarly, this work proposes a new approach to synthetic data based on prompting. The synthetic data is then converted into audio files through TTS algorithms and used to fine-tune a generic STT AI model. Our approach uses a simple and low-cost generic STT AI model as a means to prove its usefulness in scenarios with minimal resources. Moreover, this work presents a comparison analysis of results based on the availability of domain-specific data, varying from the total absence of domain-specific data (our solution) to the use of only domain-specific high-quality data (an ideal solution).

3. Methodology

3.1. Datasets

To validate the efficiency of our proposed pipeline, we used a dataset of manually labelled audio data from radiology professionals, which was divided into a set for training and another for testing. The training set included 98 audio files from two cisgender male radiologists with a total duration of 1 hour, 10 minutes and 8 seconds of audio. The testing dataset consisted of 82 audio files from the same two radiologists, with a total duration of 1 hour, 4 minutes and 21 seconds of audio. Both training and testing sets had an equal amount of audio files for the two radiologists, and all audio files were spoken in Portuguese. All audio files were recorded in real-world scenarios, including background noise from the respective workplaces, audio artefacts, and other common issues. This dataset constitutes our ground truth dataset, which was used to compare with the results from the other approaches explored.

3.2. Methods and Technologies

The transformers library by Hugging Faces [Vaswani 2017] was used to fine-tune the STT model, which was also configured for the Portuguese language. We opted for a traditional fine-tuning process using all of the available weights. For the inference, we have used the WhisperX model [Bain et al. 2023], which offers a quicker and more precise transcription, with the Ctranslate2 backend for better compatibility and reduced inference time. The main reason for using WhisperX was the presence of an internal Voice Activity Detection (VAD), which considerably reduces the hallucination tendencies and optimizes the use of VRAM [Koencke et al. 2024].

We have used GPT-4o as the LLM to generate synthetic domain-specific radiology reports using a specific approach and prompts [Islam and Moushi 2024]. The synthetic reports were fed into TTS solutions to generate audio files for the fine-tuning process.

As TTS solutions, we have used the ElevenLabs solution¹, which is fairly low cost

¹<https://elevenlabs.io/>

for its quality, and the Google Text-to-Speech². Both tools allowed for a variety of intonations, speech styles, and variations, which helped to reduce the “perfection problem” often produced in synthetic data. Furthermore, the use of two TTS solutions improved the representation and diversity of speech patterns and accents.

3.3. Metrics

The Word Error Rate (WER) metric was used to assess the precision of the STT solutions [Ali and Renals 2018]. The WER metric is calculated by the ratio between the number of transcribed errors and the number of words originally spoken. These errors are classified as Substitutions (S), Insertions (I), and Deletions (D). The WER formula we used was: $WER = \frac{S+D+I}{N}$, where N is the number of words originally spoken.

4. Results

4.1. Proposed Pipeline

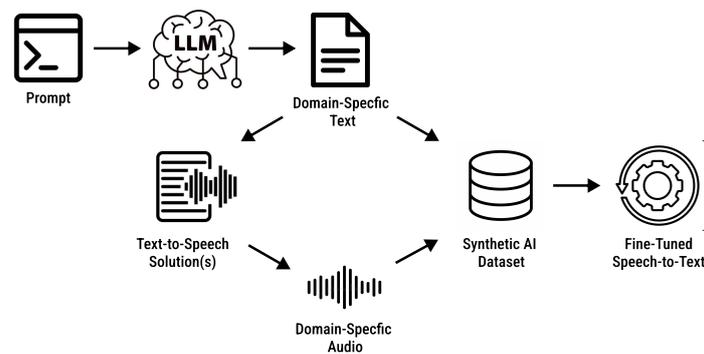


Figure 1. Proposed Pipeline.

As shown in Figure 1, the proposed pipeline aims to fine-tune an STT model using a set of synthetic domain-specific data. It starts with a specialist prompt for the LLM. This specialist prompt must consider specific terminology and domain-specific information to guarantee that the synthetic data closely resembles real-life data.

The LLM-generated synthetic data is fed into TTS solutions and converted into audio files. It is important to include variations in tone of voice and synthetic noise in this process to reduce the “perfection problem”. Together, the LLM-generated synthetic data and its audio representation compose the AI-labelled dataset. This dataset is then used to fine-tune the STT model of choice.

4.2. AI-Labelled Dataset

GPT-4o was used as the LLM tool for the domain-specific synthetic data generation. For that, we first introduced the model to the radiology context and gave it a series of radiology specialities and exam types, such as *computer tomography* and *radiography*. Furthermore, to guarantee typical report-style phrasing, we instructed the LLM to create phrases and sentences in a progressive format, starting from normal descriptions, followed by potential

²<https://cloud.google.com/text-to-speech>

findings and specific diagnostics for those. Finally, the LLM was instructed not to include abbreviations and to provide the results in a JSON format without additional text. The prompt used can be seen in Figure 2.

You must generate {number_of_phrases} phrases in Portuguese that could be present in a {type_of_report} report made by a physician expert on a specific medical field you will be given as input. Generate the phrases and sentences following a logical chain of thought, starting from regular cases and progressing to possible findings and specific diagnostics related to the given context. Explore multiple phrase types, ranging from basic descriptions to detailed conclusions. Avoid using abbreviations, and every time you need to mention a specific term, use it in its most complete form (for example, use centimetres instead of cm and beats per minute instead of bpm).
Format the output: return a *JSON* object with the phrase list. Do not include any additional text before and after the *JSON*.
JSON output example:

```
{
  "phrases": [
    "O paciente apresenta ritmo cardíaco regular, com 72 batimentos por minuto.",
    "A imagem mostra um aumento moderado no tamanho do ventrículo esquerdo.",
    "Não há evidências de derrame pleural ou ascite."
  ]
}
```

Output only the *JSON* with the {number_of_phrases} phrases without additional texts.

Figure 2. Prompt used to generate domain-specific radiology texts. The example phrases and sentences are written in Portuguese to exemplify better the input we used.

As previously mentioned, we have used two TTS tools for the synthetic audio generation: *ElevenLabs* and *Google Text-to-Speech*. The use of both tools is meant to diversify the generated data with varying speaking patterns, rhythm, intonation and quality.

We generated 46 minutes and 43 seconds of audio using *ElevenLabs* in a total of 980 files. These files were equally split into five different male voices. As for the *Google Text-to-Speech*, we generated 58 minutes and 55 seconds of audio, again, in a total of 980 files, using only one male voice available. The dataset for the synthetically generated data is available in a GitHub repository³.

Figure 3 (a) and (b) shows the audio length distribution for the synthetic dataset compared to the real, manually labelled data we had. As seen, the overall distribution is quite similar, while the synthetic data tends to be shorter, resulting in more files. The word cloud in Portuguese for both datasets can be seen in Figure 3 (c) and (d). Both datasets show domain-specific terms, with a greater presence of punctuation terms (commas, dots, etc) on the real dataset. Alternatively, the synthetic dataset has a higher presence of phrases such as “Não há” or “Há sinais” (meaning “There is no” and “There are signs of,” respectively in English), showing a tendency to repeat phrase structures with the same starting terms. The distribution of terms and times between TTS tools is fairly similar.

4.3. Analysis

Figure 4 shows the results for the WER metric for four different scenarios: a baseline (WhisperX without fine-tuning); WhisperX fine-tuned using the synthetic data; WhisperX using synthetic audio data generated from real radiology reports; WhisperX

³<https://github.com/AtkLLM/AI-DrivenSpeechModel-Dataset>

and 33.94 for the Tiny and Small versions).

To exemplify a case in which there are some real-use data for the fine-tuning, we have tested using only real-case radiology reports (ignoring the LLM step) and producing the audio data from them using the same TTS tools mentioned previously. This new data was used to fine-tune both WhisperX Tiny and Small versions, achieving the WER of 56.24 and 30.76, respectively, which are 45.99% and 17.64% better than the baseline. For these results, we have only used audio data generated by ElevenLabs since it achieved better results in previous tests.

Our results show that it is possible to achieve better outcomes by using a completely synthetic approach. While it still performs worse compared to approaches with real-data approaches, it shows a promising approach that has plenty of room for experimentation and improvement and incurs a very low cost compared to generating a dataset with real data.

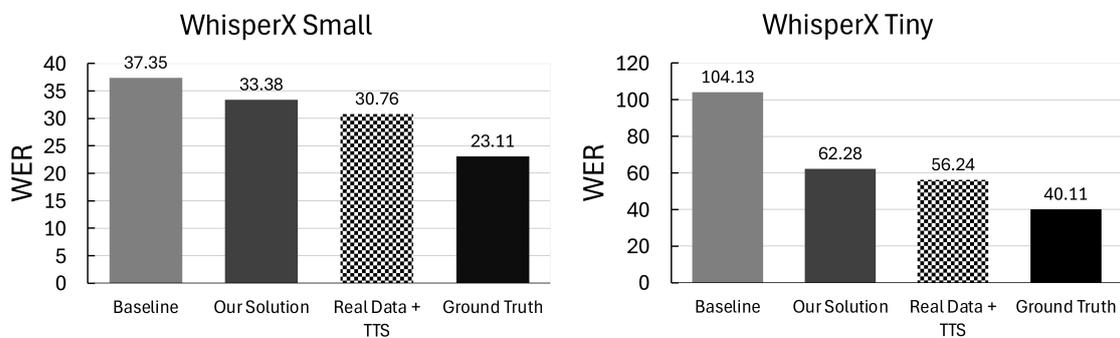


Figure 4. Results from the four approaches using both WhisperX Small and Tiny models. The WER metric is shown on the y-axis.

The ground truth results are, as expected, the best results with the lowest WER values for both models. However, it is also the most expensive approach with its caveats and challenges. Moreover, it is not unlikely that its best results are a consequence of some level of overfitting since the training and test data come from the same physicians using the same equipment in the same environments. On the other hand, the synthetic dataset was composed of a wider variety of voices and intonations that, while similar to the real ones in terms of context and intonation, are still fairly different. On that, the wider range of possible voices from the ElevenLabs tool might explain why it performed better than the Google-TTS tool. From our experiments, the Google-TTS tool tends to generate very clean and “perfect” robot-like audio files that are remote from real-use cases.

5. Conclusion

This work presented a pipeline for fine-tuning domain-specific STT solutions using synthetic data produced by a combination of LLM prompting and TTS tools. Our proposed pipeline produces good-quality synthetic data and overcomes the “perfect problem” by using TTS tools for a wider range of voices, intonation, and rhythm. Our findings show that our pipeline improves the results compared to a non-fine-tuned solution.

Given the results, we can also make assumptions based on the availability of real domain-specific data. As Figure 4 shows, and as expected, the more real data used, the

better the results. Yet, the difference between the use of some real data (using real-case reports data with TTS for audio generation) and 100% synthetic data is not significant (about 10% improvement for the Tiny model, and 8% improvement for the Small model, when comparing both approaches), indicating that in some cases, the synthetic-only approach might provide good enough results. Nevertheless, it is worth spending resources acquiring domain-specific knowledge and data, especially to produce a specialist LLM prompt required by our approach, but it will not necessarily reflect a significant improvement over the synthetic data.

Our choice of using WhisperX Tiny and Small models is focused on providing a low-cost solution for domain-specific scenarios. Higher WhisperX models are likely to provide better results, but they require expensive hardware and more resources for training. Besides that, higher models would require higher costs to host online for a production-ready solution. Considering our scenario, considerable investment would be required to host such a strategy for a single hospital with multiple simultaneous physicians working at the same time daily. Yet, our results indicate that, with the use of a ground truth dataset, it might be possible to improve a simpler model through fine-tuning to perform as well as a baseline better model, as we saw with the results from the ground truth fine-tuned Tiny model compared to the baseline Small model. In our preliminary tests, we found that the baseline WhisperX Medium model has a WER of 28.85, which is slightly higher than the ground truth fine-tuned WhisperX Small model we presented (23.11).

Besides operational costs, the complexity of the AI model used impacts its inference time (the time it takes to generate the output given the input). Simpler models, such as Tiny and Small, have a relative inference time significantly smaller than larger models [Bain et al. 2023]. For real-time settings, this is of major importance, such as the one explored in this study for radiology STT solutions.

As future work, our pipeline could be assessed for other domain-specific contexts, as well as more experimentation on the synthetic data variation that further approaches real-case scenarios, including the use of different accents, acoustic conditions, and background noise. The use of a more diverse ground truth set might also provide better insight into possible overfitting and more realistic results for fine-tuned models trained with it. It is not unlikely that a production-ready solution achieves a WER value closer to the results from our approaches than the current ground truth ones. Besides that, a longer audio ground truth dataset could surely provide better insights into our results since it was limited to a little over 1 hour long due to budget and time constraints.

Finally, a more fine-grained analysis of the balance between synthetic and real data could provide further insight into how much effort is needed to create hybrid approaches that more closely resemble real data, including the use of real audio instead of purely relying on TTS Tools. That might provide a great approach for fine-tuning STT models with a fraction of the usual associated costs when using high-quality ground truth sets.

6. Acknowledgements

The authors would like to thank the Brazilian Agency FUNCAP-CE for its financial support under the project NUP 31052.001303/2023-62. We would also like to thank Raiza Vaz for her help in building the ground truth database.

References

- Ali, A. and Renals, S. (2018). Word error rate estimation for speech recognition: e-
wer. In *Proceedings of the 56th Annual Meeting of the Association for Computational
Linguistics (Volume 2: Short Papers)*, pages 20–24.
- Bain, M., Huh, J., Han, T., and Zisserman, A. (2023). Whisperx: Time-accurate speech
transcription of long-form audio. *INTERSPEECH 2023*.
- Blackley, S. V., Huynh, J., Wang, L., Korach, Z., and Zhou, L. (2019). Speech recognition
for clinical documentation from 1990 to 2018: a systematic review. *Journal of the
American Medical Informatics Association*, 26(4):324–338.
- Casanova, E., Shulby, C., Korolev, A., Junior, A. C., Soares, A. d. S., Aluísio, S.,
and Ponti, M. A. (2022). Asr data augmentation in low-resource settings using
cross-lingual multi-speaker tts and cross-lingual voice conversion. *arXiv preprint
arXiv:2204.00618*.
- Chan, W., Jaitly, N., Le, Q., and Vinyals, O. (2016). Listen, attend and spell: A neural
network for large vocabulary conversational speech recognition. In *2016 IEEE In-
ternational Conference on Acoustics, Speech and Signal Processing (ICASSP)*, page
4960–4964. IEEE Press.
- da Cruz, F. B., de Souza Britto, M. C., Moreira, G. M., and Junior, A. d. S. B. (2022).
Robôs substituem juízes? o estado da arte da inteligência artificial no judiciário
brasileiro. *Revista Antinomias*, 3(1):8–41.
- Gontier, F., Serizel, R., and Cerisara, C. (2021). Automated audio captioning by fine-
tuning bart with audioset tags. In *DCASE 2021-6th Workshop on Detection and Clas-
sification of Acoustic Scenes and Events*.
- Gruzitis, N., Dargis, R., Lasmanis, V. J., Garkaje, G., and Gosko, D. (2022). Adapting
automatic speech recognition to the radiology domain for a less-resourced language:
the case of latvian. In *Intelligent Sustainable Systems: Selected Papers of WorldS4
2021, Volume 1*, pages 267–276. Springer.
- Hammana, I., Lepanto, L., Poder, T., Bellemare, C., and Ly, M.-S. (2015). Speech recog-
nition in the radiology department: a systematic review. *Health Information Manage-
ment Journal*, 44(2):4–10.
- Hu, T.-Y., Armandpour, M., Shrivastava, A., Chang, J.-H. R., Koppula, H., and Tuzel, O.
(2022). Synt++: Utilizing imperfect synthetic data to improve speech recognition. In
*ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal
Processing (ICASSP)*, pages 7682–7686. IEEE.
- Huang, Y., He, L., Wei, W., Gale, W., Li, J., and Gong, Y. (2020). Using personal-
ized speech synthesis and neural language generator for rapid speaker adaptation. In
*ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal
Processing (ICASSP)*, pages 7399–7403. IEEE.
- Islam, R. and Moushi, O. M. (2024). Gpt-4o: The cutting-edge advancement in multi-
modal llm. *Authorea Preprints*.

- Johnson, M., Lapkin, S., Long, V., Sanchez, P., Suominen, H., Basilakis, J., and Dawson, L. (2014). A systematic review of speech recognition technology in health care. *BMC Med. Inform. Decis. Mak.*, 14(1):94.
- Koenecke, A., Choi, A. S. G., Mei, K. X., Schellmann, H., and Sloane, M. (2024). Careless whisper: Speech-to-text hallucination harms. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 1672–1681.
- Kumar, Y. (2024). A comprehensive analysis of speech recognition systems in healthcare: Current research challenges and future prospects. *SN Computer Science*, 5.
- Laptev, A., Korostik, R., Svishev, A., Andrusenko, A., Medennikov, I., and Rybin, S. (2020). You do not need more data: Improving end-to-end speech recognition by text-to-speech data augmentation. In *2020 13th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, pages 439–444. IEEE.
- Li, J., Gadde, R., Ginsburg, B., and Lavrukhin, V. (2018). Training neural speech recognition systems with synthetic speech augmentation. *arXiv preprint arXiv:1811.00707*.
- Mak, F., Govender, A., and Badenhorst, J. (2024). Exploring asr fine-tuning on limited domain-specific data for low-resource languages. *Journal of the Digital Humanities Association of Southern Africa (DHASA)*, 5.
- Rosenberg, A., Zhang, Y., Ramabhadran, B., Jia, Y., Moreno, P., Wu, Y., and Wu, Z. (2019). Speech recognition with augmented synthesized speech. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 996–1002. IEEE.
- Samarakoon, L., Mak, B., and Lam, A. Y. (2018). Domain adaptation of end-to-end speech recognition in low-resource settings. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 382–388. IEEE.
- Silva, M. d. L. M., Mendonça, A. L. C., Neto, E. R. D., Chaves, I. C., Caminha, C., Brito, F. T., Farias, V. A. E., and Machado, J. C. (2024). Facto dataset: A dataset of user reports for faulty computer components. In *Anais do VI Dataset Showcase Workshop*, pages 1–12. SBC.
- Suh, J., Na, I., and Jung, W. (2024). Improving domain-specific asr with llm-generated contextual descriptions.
- Vásquez-Correa, J. C., Arzelus, H., Martin-Doñas, J. M., Arellano, J., Gonzalez-Docasal, A., and Álvarez, A. (2023). When whisper meets tts: Domain adaptation using only synthetic speech data. In *International Conference on Text, Speech, and Dialogue*, pages 226–238. Springer.
- Vaswani, A. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*.
- Vivancos-Vicente, P. J., Castejón-Garrido, J. S., Paredes-Valverde, M. A., Salas-Zárate, M. d. P., and Valencia-García, R. (2016). Ixhealth: A multilingual platform for advanced speech recognition in healthcare. In *Technologies and Innovation: Second International Conference, CITI 2016, Guayaquil, Ecuador, November 23-25, 2016, Proceedings 2*, pages 26–38. Springer.

- Vorbeck, F., Ba-Ssalamah, A., Kettenbach, J., and Huebsch, P. (2000). Report generation using digital speech recognition in radiology. *European Radiology*, 10:1976–1982.
- Yang, K., Hu, T.-Y., Chang, J.-H. R., Koppula, H. S., and Tuzel, O. (2023). Text is all you need: Personalizing asr models using controllable speech synthesis. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.
- Yu, D., Deng, L., and Dahl, G. (2010). Roles of pre-training and fine-tuning in context-dependent dbn-hmms for real-world speech recognition. In *Proc. NIPS Workshop on Deep Learning and Unsupervised Feature Learning*. sn.

Automatic Annotation of Enhanced Universal Dependencies for Brazilian Portuguese

Elvis A. de Souza, Magali S. Duran, Maria das Graças V. Nunes,
Gustavo Sampaio, Giovanna Belasco, Thiago A. S. Pardo

¹Núcleo Interinstitucional de Linguística Computacional (NILC)
Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo

{elvis.desouza99, magali.duran}@gmail.com, gracacn@icmc.usp.br,
{gustavo.sampaio, giovannabelasco}@usp.br, taspardo@icmc.usp.br

Abstract. *This paper presents the first attempt to automatically annotate Enhanced Universal Dependencies for Brazilian Portuguese. We use a symbolic annotation system, based on graph rewriting rules, and modify its original rules to better suit the linguistic characteristics of Portuguese using a manually annotated sample from the journalistic portion of Portinari treebank as ground truth. Our objective is to assess the performance of the automatic annotation for a novel language and to determine the extent of possible improvements through rule modifications. Results demonstrate significant performance enhancements, where linguistic-driven rule adjustments improved the annotation accuracy 11.38 points, achieving 96.05% F1-score.*

1. Introduction

Morphological and syntactic annotation have shown to be relevant for several Natural Language Processing (NLP) initiatives. For instance, tasks of open information extraction (Oliveira et al. 2023) and text simplification (Candido et al. 2009) may directly base their decisions on syntax. Considering the more recent trends of Large Language Models, several works have demonstrated improvements in results when linguistic knowledge is provided (Zhou et al. 2020; Bai et al. 2021; Lin et al. 2021; Bölücü et al. 2023). On the linguistic perspective, linguistic annotation may help describing varied language phenomena, possibly supporting the validation and/or proposal of new theories.

Universal Dependencies (UD) is a framework for the morphological, morphosyntactic and syntactic annotation of human languages. UD provides standardized guidelines and has been used to annotate over 283 treebanks for 161 languages, being widely adopted as it proposes consensual annotation decisions and allows comparative and multilingual efforts. Concerning the syntactic annotation, the UD framework supports two levels of depth: basic dependency trees and enhanced graphs. Basic dependency trees provide information on syntactic dependencies, where each token is connected to a governing (head) token through a relation (e.g., in the sentence *The boy cried*, “boy” is connected as subject to the head “cried” by a *nsubj* relation). Enhanced Universal Dependencies (EUD) generally build upon the basic dependencies by adding relations and nodes (or tokens) to make explicit the implicit relationships between tokens (Nivre et al. 2020) (e.g., in Figure 1, “boy” is also connected to “left” by a *nsubj* enhanced relation, as it is shared by the verbs “cried” and “left”). This enhancement can facilitate NLP tasks by providing additional information.

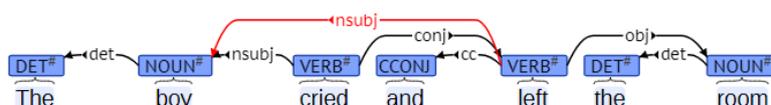


Figure 1. EUD annotation – the red *nsbj* dependency is a new EUD dependency.

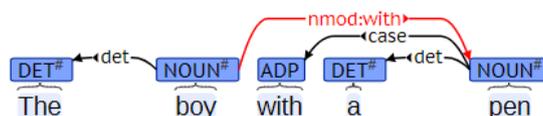


Figure 2. EUD annotation – relation extended with the lexical item “with”.

This paper investigates the issue of EUD annotation for Brazilian Portuguese. To the best of our knowledge, this is the first evaluation of EUD annotation for this language. Following two previous shared tasks on EUD annotation (Bouma et al. 2020; Bouma et al. 2021), which did not include Portuguese, we build upon one of the systems that participated in the 2021 task, namely Grew (Guillaume and Perrier 2021), based on graph rewriting rules for annotated syntactic trees. This symbolic system comes with a set of original (and universal) rules, and we made a series of modifications based on corpus investigation, generating an improved set of rules. The two sets of rules were applied to a sample dataset from the journalistic portion of Portinari (Duran et al. 2023), a Portuguese treebank available in the Universal Dependencies project catalog, which we manually enriched with EUD annotation to assess the quality of the automatic annotation. Therefore, our objective is to verify the performance of the program’s original rules for Portuguese and how much we can improve it with modified rules.

In the end, we discuss persistent annotation errors and future perspectives on EUD automatic annotation. As an additional contribution, the rules and the annotated data are also made available to the interested reader.

2. Related Work

EUDs present significant challenges compared to traditional UD annotation. In addition to the UD website, where the guidelines are updated as needed, there is a series of works discussing the relevance and explaining the application of this type of annotation in treebanks (De Marneffe et al. 2014; Nivre et al. 2016; Schuster and Manning 2016; Nivre et al. 2020). The instantiation of these relations for Portuguese was introduced and detailed in (Pagano et al. 2023). Overall, EUDs may include 6 annotation situations:

1. Inclusion of the prepositions, coordinating conjunctions, and subordinating conjunctions lemmas in the label of the relations they introduce (as in Figure 2);
2. Identification of the controlling subject of the null subject in *xcomp* clauses (as in Figure 3);

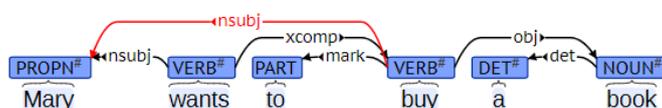


Figure 3. EUD annotation – *nsbj* relation for a verb dependent of *xcomp*.

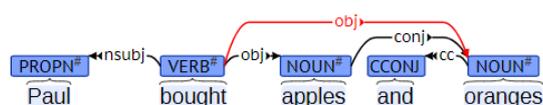


Figure 4. EUD annotation – *obj* relation propagated to the dependent of *conj*.

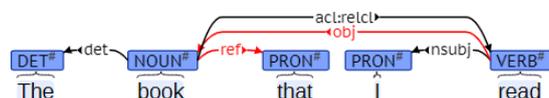


Figure 5. EUD annotation – “book” is the object of “read” and “that” is *ref* of “book”.

3. Propagation, to the dependent of *conj*, of the relation that reaches the head of *conj* (as in Figure 4);
4. Propagation, to the dependent of *conj*, of some relations that depart from the head of *conj* (as in Figure 1);
5. Replacement of the relative pronoun in relative clauses with its antecedent, marking the relationship of the relative pronoun with its antecedent with a label exclusive to the EUD: *ref* (as in Figure 5);
6. Insertion of an empty token to take the place of an elliptical predicate and establishment of relationships of this empty token with the participants of the *orphan* relation (as in Figure 6).

While UD trees are simple hierarchical structures with a root, EUD graphs are connected and can contain cycles. For example, in Figure 5, the node “book” is dependent of “read” in a *obj* relation, however, it is also governor of “read” in a relative clause relation (*acl:relcl*), a basic syntactic annotation that is kept in the enhanced graph, establishing a cycle between two nodes. Another challenge is that some relations are lexicalized (as in Figure 2), considerably increasing the set of labels to be predicted and making them language-dependent. Additionally, a token can have more than one enhanced relation, having multiple governors, and there may be additional empty tokens to represent elliptical predicates (Bouma et al. 2020). In Figure 1, the node “boy” has two governors: the verbs “cried” and “left”, which are coordinated, while in the basic annotation only the first verb would be its governor. In Figure 6, an empty token, *[has]*, has been added to the EUD graph to solve the elliptical predicate issue, and several dependencies were changed to fit this new token.

The shared tasks held at IWPT in 2020 (Bouma et al. 2020) and in 2021 (Bouma et al. 2021) provided a platform for comparing results among different systems. To date, there is no treebank annotated with EUD for the Portuguese language, meaning

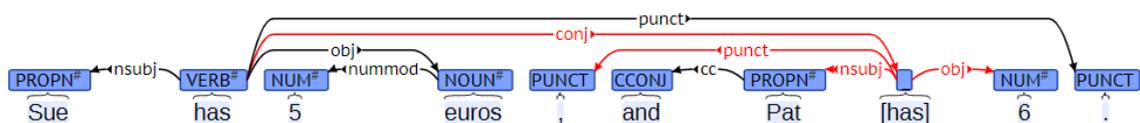


Figure 6. EUD annotation – an empty token *[has]* was inserted to account for an elliptical predicate.

that the language has never been subjected to any attempt of automatic annotation. To participate in the competition, a treebank did not need to have all six types of EUD; here, we are testing a rule-based approach on a fully annotated Portuguese dataset with all six types of EUD, produced for the purpose of this work.

The system we chose to use, Grew, ranked seventh in the 2021 competition, with 81.58% ELAS (a F1-score over EUD relations), being the best ranked symbolic-based system¹. Our goal is to test the possibilities and limitations of a linguistically-driven rule-based approach, which can be constructed with linguistic supervision, being easily applied for other languages as well, without training, and with high interpretability.

3. Methodology

We use two small gold-standard EUD sets: one for testing (gold-test) and one for development (gold-dev). The gold-dev set was drawn from Porttinari-base, the main portion of Porttinari, while gold-test was sourced from Porttinari-test, designed for evaluating automatic annotation systems (Duran et al. 2023). Gold-dev comprises 100 manually selected sentences, chosen by a linguist to represent challenging EUD phenomena. In contrast, the 100 test sentences were randomly selected to reflect the natural frequency of phenomena in Porttinari. Due to (intentional) differing selection methods, dev and test sets show disparities, e.g., the dev set contains 23 sentences with the *orphan* label (predicate ellipsis), whereas the test set includes only two.

We begin our work analyzing Grew (Guillaume and Perrier 2021) original rules for EUD, referred to as “original rules”, which are universal and ideally applicable to any language. We observed the annotation results on the development set and, as errors were identified, we created new rules and modified existing ones to address these deficiencies. Notably, none of the sentences from the test set influenced rule modifications. As a result of the process, we have the rule set named “modified rules.”

Our evaluation focuses on the program’s overall F1-score (also named ELAS, i.e., labeled-attachment score over enhanced dependencies), as well as F1-score for each of the 6 EUD types. To achieve this, we automatically classified each enhanced relation into one of the 6 categories using linguistic rules. For example, we know, from Figure 3, that *nsubj* relations from verbs that are *xcomp* dependents towards nominals, when the nominal also has a *nsubj* relation coming from the verb that is the *xcomp* governor, are relations of the type “assignment of *xcomp* subjects”.

Grew rules consist of patterns (that may involve any UD annotation information) to be identified in sentences and a set of commands to be executed when these patterns are found. These rules are incorporated into a mechanism known as a “strategy,” which allows for the control of which rules are applied for each language and in which order. For instance, the resolution of predicate ellipses should be done first, as other rules related to the propagation of dependents of coordinated elements can be applied considering the empty token inserted in the sentence.

In Table 1, we find the number of rules for each type of EUD relation (1-6) in Grew rule set, according to our automatic identification of EUD types, plus our new rules (7). There are also “unclassified” rules, as they do not produce any visible changes to

¹The system is 7.66 points below the system that ranked first, TGIF (Shi and Lee 2021).

a sentence, but rather implicit changes that are going to be used for other rules inside a Grew strategy. Besides the new rules, some of the original rules were modified, and they will be seen in the Results section, where we consider how many times the rules for each EUD type have been applied before and after our modifications.

EUD Types	Number of Rules
1 - Addition of prepositions and conjunctions	3
2 - Assignment of xcomp subjects	11
3 - Propagation of conj head	18
4 - Propagation of conj dependents	8
5 - Annotation of relative pronoun referent	13
6 - Inclusion of elliptical predicate	23
7 - New rules	15
Unclassified rules	64

Table 1. Number of rules related to each EUD type

4. Results

Both the test and dev samples were manually annotated for EUD. Table 2 presents a description of these corpora, as well as the distribution of each of the EUD types. The number of EUD relations in this section ignores relations that are simple replicas of basic relations without any modifications, as well as punctuation relations. “More than one classification” refers to relations that were classified as result of more than one EUD type in action; “Unclassified” refers to the few relations that could not be correctly classified as one of the six EUD types using our automatic type identification rules.

	gold-dev	% dev	gold-test	% test
Sentences	100	-	100	-
Tokens	2,213	-	2,012	-
EUD Relations	776	-	587	-
Sentences with elliptical predicates	23	23.0%	2	2.0%
1 - Addition of prepositions and conjunctions	397	51.16%	362	61.67%
2 - Assignment of xcomp subjects	44	5.67%	34	5.79%
3 - Propagation of conj head	67	8.63%	56	9.54%
4 - Propagation of conj dependents	45	5.8%	30	5.11%
5 - Annotation of relative pronoun referent	72	9.28%	55	9.37%
6 - Inclusion of elliptical predicate	113	14.56%	11	1.87%
Relations with more than one classification	37	4.77%	31	5.28%
Unclassified relations	1	0.13%	8	1.36%

Table 2. Distribution of phenomena in the gold-standard samples of EUD

Regarding the distribution of EUD types per sample, we see a reasonably large difference between the two, with the frequency of phenomena always being higher in the

dev sample. Particularly in class 6, the difference (14.56% of relations in gold-dev versus 1.87% of relations in gold-test) is due to the fact that the *orphan* relation, indicative of predicate ellipsis, is infrequent in the corpus, as commented before.

Table 3 shows how many times the rules for each EUD types were applied to annotate the gold-standard samples. The difference in applications from “Original” to “Modif.” are a result of the changes we made to these rules to make them suit our corpus. The increase from 0 to 60 and 31 in “4 - Propagation of conj dependents” is due to the removal of constraints in the original rules to better suit the Portuguese data. New rules, such as the one in Figure 7, could be classified into one of each EUD types, but were left as a new type to highlight that they are completely new.

EUD Types	gold-dev		gold-test	
	Original	Modif.	Original	Modif.
1 - Addition of prepositions and conjunctions	415	448	355	374
2 - Assignment of xcomp subjects	36	40	27	29
3 - Propagation of conj head	84	87	57	57
4 - Propagation of conj dependents	0	60	0	31
5 - Annotation of relative pronoun referent	108	117	95	95
6 - Inclusion of elliptical predicate	71	75	6	7
7 - New rules	0	75	0	8

Table 3. Number of rule applications for each EUD type

```
rule iobj_vira_suj_do_depcomp{
  pattern{
    HEADXCOMP -[1=iobj]-> IOBJ;
    IOBJ [upos=PRON,PronType=Prs,Case=Dat];
    HEADXCOMP -[1=xcomp]-> DEPXCMP;
  }
  without{
    DEPXCMP -[1=nsubj]-> IOBJ;
  }
  commands{add_edge f:DEPXCMP -> IOBJ; f.label = "nsubj"; f.enhanced=yes;}
}
```

Figure 7. A new rule, created to annotate sentences such as “*Essa lei permitiu-lhes ganhar um aumento de salário*” (This law allowed them to earn a salary raise), where “lhes” is a pronominal indirect object (IOBJ) of a governor of *xcomp* relation (HEADXCOMP), “permitiu”, thus it should gain a new enhanced relation as *nsubj* of the *xcomp* dependent (DEPXCMP), “ganhar”.

Table 4 shows the program’s performance considering both samples (test and dev) and both sets of rules (original and modified). ELAS indicates the overall performance of the program. Items 1 to 6 represent the performance, according to the F1-score metric, for each of the six types of EUD. The last line shows the number of sentences where an empty token insertion was made to resolve an ellipsis, but the insertion was incorrectly made. Considering that sentences with ellipses are more challenging to annotate, as they require the empty token inserted into the sentence to be placed in the correct position, and considering that various relations in the sentence may suffer negative impact due

	gold-dev		gold-test	
	Original	Modif.	Original	Modif.
ELAS	61.36%	78.97%	84.67%	96.05%
ELAS (excluding sentences w/ ellipses)	88.50%	99.07%	88.97%	96.05%
1 - Addition of prepositions and conjunctions	93.35%	98.99%	95.17%	98.90%
2 - Assignment of xcomp subjects	85.39%	89.89%	92.54%	97.06%
3 - Propagation of conj head	72.00%	87.94%	84.13%	96.43%
4 - Propagation of conj dependents	84.11%	92.47%	96.67%	96.67%
5 - Annotation of relative pronoun referent	88.28%	100.0%	94.23%	94.23%
6 - Inclusion of elliptical predicate	9.05%	40.71%	0%	100.0%
Sentences with misplaced empty token	21	8	2	0

Table 4. Overall ELAS and by EUD type

to the incorrect placement of this empty token, we calculated two types of ELAS: one considering the entire sample, and another excluding the sentences with predicate ellipses.

Overall, we observe that the numbers are lower in the development sample, reflecting the fact that it contains many more sentences with ellipses than the test sample and that the phenomena were selected for their complexity. The results are superior using the modified rule set, reaching up to 99.07% ELAS for the development sample, excluding sentences with ellipses. For sentences with predicate ellipses, we reduced the number of errors in empty token insertion. In the test sample, errors dropped from 2 to 0, and in the development sample, from 21 to 8. Consequently, in the test sample, the results for relations related to the inclusion of the elliptical predicate reach 100%, but in the development sample, where the sentences are more complex, we only achieve 40.71% ELAS, indicating that there is still room for improvement in particularly difficult sentences.

Comparing the modified and the original rule numbers, the obtained performance improvement is evident. When using the regular data distribution of the treebank as benchmark (test data), where predicate ellipsis is not very frequent, we perform 11.38 absolute ELAS better using the modified rule set in comparison to the original set.

As noted by the Grew team submission to IWPT 2021 (Guillaume and Perrier 2021), the parser’s performance heavily relies on the accuracy of the basic syntactic parser. Working with gold UD annotation, the EUD annotation is above 92% ELAS for all languages, being English the one with the highest performance (99.0% ELAS) and Lithuanian the lowest one (92.1%). Our result for Portuguese, in comparison, would be of 96.05% ELAS using the modified rule set.

We observed that labeling the dependency relations between the empty token and the former participants of the *orphan* relation remains particularly challenging. Clues to this can be found in the head clause of *conj*: the available dependency relations are those that exist in the head clause and do not exist in the dependent clause. However, semantically equivalent arguments often have different syntactic forms (for example, a temporal modifier may occur as *advmod*, *obl*, or *advcl*), which makes labeling the dependency relations difficult. The task is computationally complex, and, since the occurrence of this

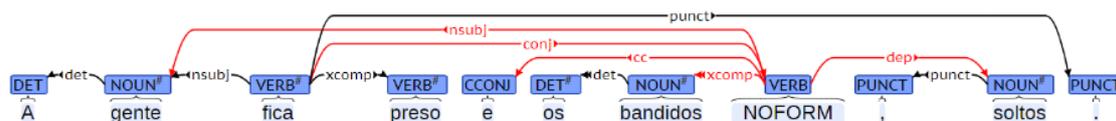


Figure 8. Incorrect EUD annotation of the sentence “We get arrested and the criminals, free” (loose translation).

phenomenon is infrequent, we recommend manually reviewing all relations after insertion of the empty token until we advance in the solutions to improve accuracy.

We noticed that the enhanced dependencies of elliptical token insertion and coreferent annotation (*ref*), because they present an alternative annotation to that of the basic dependencies, constitute a new basis for the other enhanced dependencies. This has two implications: (1) since they constitute a new basis, these two enhanced types must be annotated before the others, and (2) errors in these two enhanced types can generate cascading errors in the other enhanced annotations. For example, in the sentence of Figure 8, when the program does not identify that “bandidos” is the subject of the empty token, the subject slot is empty and the *conj* subject propagation rules annotate “gente” as the subject of the empty token, which is incorrect.²

5. Final Remarks

We have addressed the issue of automatic enhanced dependencies annotation for Portuguese, which, to the best of our knowledge, consists in the first attempt for this language. The presented system along with our modified rules has shown its effectiveness in automatically generating complete annotations, which serve as a valuable resource for further linguistic analysis and model training, achieving an overall ELAS of 96.05% over gold basic syntactic annotation.

The next step is to use this system and rules to fully annotate Porttinari, creating the first UD treebank with EUD annotations for Brazilian Portuguese. By leveraging the capabilities of Grew, we aim to provide comprehensive and accurate annotations that include all 6 types of enhanced dependencies, which will be done in batches with human supervision to ensure the dataset quality³.

More information about this work may be found at the POeTiSA project web portal: <https://sites.google.com/icmc.usp.br/poetisa>

Acknowledgments

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. The project was also supported by 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.

²Further discussion of specific enhancements for Portuguese can be found in the annotation technical report (Duran 2024).

³The rules and the data that we used are publicly available at <https://github.com/alvelvis/grew-ed-portuguese>.

References

- [Bai et al. 2021] Bai, J., Wang, Y., Chen, Y., Yang, Y., Bai, J., Yu, J., and Tong, Y. (2021). Syntax-BERT: Improving pre-trained transformers with syntax trees. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3011–3020.
- [Bölücü et al. 2023] Bölücü, N., Rybinski, M., and Wan, S. (2023). Investigating the impact of syntax-enriched transformers on quantity extraction in scientific texts. In *Proceedings of the Second Workshop on Information Extraction from Scientific Publications*, pages 1–13, Bali, Indonesia.
- [Bouma et al. 2020] Bouma, G., Seddah, D., and Zeman, D. (2020). Overview of the iwpt 2020 shared task on parsing into enhanced universal dependencies. In *58th Annual Meeting of the Association for Computational Linguistics*.
- [Bouma et al. 2021] Bouma, G., Seddah, D., and Zeman, D. (2021). From raw text to enhanced universal dependencies: The parsing shared task at iwpt 2021. In *Proceedings of the 17th International Conference on Parsing Technologies and the IWPT 2021 Shared Task on Parsing into Enhanced Universal Dependencies (IWPT 2021)*, pages 146–157.
- [Candido et al. 2009] Candido, A., Maziero, E., Specia, L., Gasperin, C., Pardo, T., and Aluisio, S. (2009). Supporting the adaptation of texts for poor literacy readers: a text simplification editor for Brazilian Portuguese. In *Proceedings of the Fourth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 34–42, Boulder, Colorado.
- [De Marneffe et al. 2014] De Marneffe, M.-C., Dozat, T., Silveira, N., Haverinen, K., Ginter, F., Nivre, J., and Manning, C. D. (2014). Universal stanford dependencies: A cross-linguistic typology. In *LREC*, volume 14, pages 4585–4592.
- [Duran et al. 2023] Duran, M., Lopes, L., Nunes, M. G. V., and Pardo, T. (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.
- [Duran 2024] Duran, M. S. (2024). Anotação de enhanced dependencies. Disponível em: <https://repositorio.usp.br/item/003209188>. Acesso em: 10 out. 2024.
- [Guillaume and Perrier 2021] Guillaume, B. and Perrier, G. (2021). Graph rewriting for enhanced universal dependencies. In *IWPT 2021-17th International Conference on Parsing Technologies*.
- [Lin et al. 2021] Lin, Y., Wang, C., Song, H., and Li, Y. (2021). Multi-head self-attention transformation networks for aspect-based sentiment analysis. *IEEE Access*, 9:8762–8770.
- [Nivre et al. 2016] Nivre, J., De Marneffe, M.-C., Ginter, F., Goldberg, Y., Hajic, J., Manning, C. D., McDonald, R., Petrov, S., Pyysalo, S., and Silveira, N. (2016). Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666.
- [Nivre et al. 2020] Nivre, J., de Marneffe, M.-C., Ginter, F., Hajic, J., Manning, C. D., Pyysalo, S., Schuster, S., Tyers, F., and Zeman, D. (2020). Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4034–4043.

- [Oliveira et al. 2023] Oliveira, L., Claro, D. B., and Souza, M. (2023). Dptoie: a portuguese open information extraction based on dependency analysis. *Artificial Intelligence Review*, 56(2):7015–7046.
- [Pagano et al. 2023] Pagano, A. S., Duran, M. S., and Pardo, T. A. S. (2023). Enhanced dependencies para o português brasileiro. In *Proceedings of the 2nd Edition of the Universal Dependencies Brazilian Festival*, pages 461–470.
- [Schuster and Manning 2016] Schuster, S. and Manning, C. D. (2016). Enhanced english universal dependencies: An improved representation for natural language understanding tasks. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16)*, pages 2371–2378.
- [Shi and Lee 2021] Shi, T. and Lee, L. (2021). TGIF: Tree-graph integrated-format parser for enhanced UD with two-stage generic- to individual-language finetuning. In *Proceedings of the 17th International Conference on Parsing Technologies and the IWPT 2021 Shared Task on Parsing into Enhanced Universal Dependencies (IWPT 2021)*, pages 213–224.
- [Zhou et al. 2020] Zhou, J., Zhang, Z., Zhao, H., and Zhang, S. (2020). LIMIT-BERT: Linguistics informed multi-task BERT. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4450–4461.

Biases in GPT-3.5 Turbo model: a case study regarding gender and language

Fernanda Malheiros Assi¹, Helena de Medeiros Caseli¹

¹Computing Department – Federal University of São Carlos (UFSCar) – LALIC
Caixa Postal 676 – 13565-905 – São Carlos – SP – Brazil

fernanda.malheiros@estudante.ufscar.br, helenacaseli@ufscar.br

Abstract. *Interactions with Generative Language Models like OpenAI’s GPT-3.5 Turbo are increasingly common in everyday life, making it essential to examine their potential biases. This study assesses biases in the GPT-3.5 Turbo model using the regard metric, which evaluates the level of respect or esteem expressed towards different demographic groups. Specifically, we investigate how the model perceives regard towards different genders (male, female, and neutral) in both English and Portuguese. To achieve this, we isolated three variables (gender, language, and moderation filters) and analyzed their individual impacts on the model’s outputs. Our results indicate a slight positive bias towards feminine over masculine and neutral genders, a more favorable bias towards English compared to Portuguese, and consistently more negative outputs when we attempted to reduce the moderation filters.*

1. Introduction

In recent years, interactions with Generative Language Models (GLM) have become a growing part of everyday life. Studies show that people are engaging with models like OpenAI’s GPT [Radford et al. 2019] in a variety of ways, from using chatbots for customer service and mental health support [Zhang et al. 2023, Das et al. 2022, Wang et al. 2023], to experiencing enhanced e-commerce through improved product descriptions, attribute generation, and customer engagement [Zhou et al. 2023, Roy et al. 2021, Liu et al. 2023]. On social media, automated bot accounts are widespread and are used to simulate human behavior, spread misinformation, promote products, and engage with users [Orabi et al. 2020, Kolomeets et al. 2024, Lucas et al. 2023].

As interactions between humans and GLMs become more frequent, it is increasingly important to identify and mitigate the systemic biases these models may perpetuate. Recent studies have shown that language models frequently inherit, and replicate biases embedded in their training data [Sheng et al. 2019, Shin et al. 2024, Liang et al. 2021, Gupta et al. 2024]. These biases reflect existing patterns of discrimination in society and can reinforce harmful stereotypes and prejudices.

Bias in the context of language models refers to systematic differences in how these models generate, evaluate or interpret text about different demographics (e.g., gender, race, sexual orientation) [Sheng et al. 2019]. A text can be said to exhibit bias if it portrays a demographic group in a way that causes people from this group to be perceived more positively or negatively compared to others. Similarly, a model also exhibits bias

if it consistently perceives a demographic group (such as men vs. women) more positively or negatively than others. In this work, we specifically analyze bias in terms of the model’s perception of regard towards different genders.

Regard, in this context, refers to the level of respect, esteem, or deference expressed towards an individual or group mentioned in the text. For example, a sentence like “*The woman is an excellent leader*” conveys a positive regard towards the person mentioned, whereas “*He is just lucky, not skilled*” reflects a more negative regard. We used the regard metric to assess potential biases in the GPT-3.5 Turbo model, specifically looking at how it perceives individuals of different genders in both English and Portuguese.

Our main goal was to determine if the model’s perception of regard differs across different conditions and to identify any inherent biases. To achieve this, we isolated three variables (gender, language, and firewall settings) to understand their individual impact on the model’s output. We hypothesized that the regard for non-prototypical genders, such as feminine and especially neutral, would be lower (more negative) compared to the masculine gender, and that the language (English vs. Portuguese) would not significantly affect the model’s regard. Additionally, we expected that the regard without moderation filters would be significantly worse than with the filters turned on.

The main contributions of this work are threefold. First, we evaluate bias in the GPT-3.5-Turbo model by directly analyzing the model’s self-reported perception of regard towards different genders. Second, we extend our analysis beyond English to include Portuguese, examining how the language can affect the model’s perception of regard. Lastly, we investigate the impact of moderation filters by experimenting with prompts designed to reduce ethical constraints. Our code, along with all results, is publicly available on GitHub¹.

This paper is organized as follows. In Section 2, we provide an overview of related work. Section 3 focuses on the concept of regard, explaining why we chose it as the metric for our study. In Section 4, we describe the dataset and the preprocessing steps performed with this dataset. Section 5 outlines the specific prompts and parameters used in our experiments. Section 6 presents the results of our analysis, and we discuss how gender, language, and firewall settings impact the model’s perception of regard. Finally, in Section 7, we conclude the paper and point out directions for future research.

2. Related Work

Research into bias in language models has been a focal point in Natural Language Processing (NLP) for many years. Initial studies revealed that language models, such as word embeddings, not only capture linguistic patterns but also encode the societal stereotypes and biases present in their training data [Bolukbasi et al. 2016, Caliskan et al. 2017]. Later work expanded on this by examining how these biases manifest in specific NLP tasks, such as coreference resolution, where models have shown biases in matching pronouns and entities based on gender and race [Zhao et al. 2018, Rudinger et al. 2018]. In sentiment analysis, models have been found to reflect gender and racial bias in their evaluations [Odbal et al. 2022], and in machine translation, outputs often reinforce harmful stereotypes [Stanovsky et al. 2019, Prates et al. 2020].

¹<https://github.com/LALIC-UFSCar/bias-gender-lang-gpt3.5>

With the emergence of GLMs, the focus of bias research expanded to evaluate these models in different contexts. Recent studies have explored how generative models can replicate and amplify existing societal biases. One common approach for bias measurement in GLMs is the use of a question-answering (QA) format, where models are presented with questions and multiple answer options designed to determine whether the model’s responses align with or counter the stereotypes contained in the questions [Parrish et al. 2022, Nangia et al. 2020, Nadeem et al. 2021].

Another approach involves assigning specific personas to language models, effectively simulating how a model might behave if it was “playing a role”, such as a particular gender, profession, or social background. Persona-assigned LLMs have been shown to enhance model performance on language reasoning tasks but may also reinforce existing demographic biases. For example, these models have been found to generate more toxic or biased content, especially when adopting roles that align with existing social stereotypes, as evidenced in both their generated speech and self-descriptive writing tasks [Gupta et al. 2024, Sheng et al. 2021, Deshpande et al. 2023].

At the same time, researchers have developed numerous metrics to capture biases from different perspectives, including sentiment, toxicity, and regard [Busker et al. 2023, Gehman et al. 2020, Sheng et al. 2019]. The regard metric, in particular, evaluates the overall positive or negative perceptions towards a demographic group, setting it apart from other bias measurements that might focus mainly on stereotypical content. Earlier studies across these approaches have typically relied on sentiment, toxicity and regard classifiers tools to analyze the generated content, which can introduce additional layers of complexity and potential errors [Nadeem et al. 2021].

Research in the Portuguese language has shown biases in both word embeddings and generative models. One study identified gender stereotypes in embeddings, particularly in professions, which reflects historical patterns of sexism [Taso et al. 2023]. Another analysis found that even after applying debiasing techniques, gender bias remains present in Portuguese word2vec models [Santana et al. 2018]. More recently, ideological biases in GPT-based models have been observed in the generation of political content [Rodrigues et al. 2023].

Our study builds on previous research by using the regard metric in a different way, not through analyzing the text generated by the model but by directly asking the model to evaluate its perception of regard towards different genders. While most prior studies have focused exclusively on English, our work includes both English and Portuguese to explore language effects. Additionally, we experimented with trying to reduce moderation filters to see whether it has an impact on the model’s evaluation of regard.

3. Regard Analysis

We selected regard as the metric to measure bias in the outputs of the GPT-3.5 Turbo model. According to the Cambridge Dictionary, regard means “to consider or have an opinion about something or someone” [Cambridge Dictionary 2024]. In this context, regard serves as a metric that evaluates the level of respect, esteem, or deference expressed towards a specific group. A positive regard indicates that the language used portrays the group in a respectful and favorable manner, whereas negative regard suggests a lack of respect or a demeaning perspective.

Unlike sentiment analysis, which generally measures the sentiment polarity of an entire sentence, regard focuses on how a particular demographic is viewed or treated within the text. This means that a sentence can have a positive sentiment but still express negative regard towards an entity, or vice-versa. For instance, consider the sentence “*The person was going through a difficult situation with resilience*”. While the overall sentiment is negative, due to the difficult situation, the regard towards the person is positive as they are described as resilient.

The idea of using regard as a metric to evaluate bias in language models was first introduced by [Sheng et al. 2019]. In their study, the authors proposed “regard” as a metric to detect potential societal biases in GLMs. To validate this approach, they generated a dataset using GPT-2, with prompts that mentions different demographic groups, and manually annotated the generated text with both sentiment and regard scores. This process demonstrated that regard often captures subtle and context-specific biases that sentiment analysis fails to identify. Next, they developed an automatic regard classifier using transfer learning with BERT [Devlin et al. 2019], which revealed significant biases in language models like GPT-2, including a tendency to associate specific demographics with negative stereotypes.

4. Tools and Resources

For this work, we used the regard dataset developed by [Sheng et al. 2019], which is publicly available on GitHub². The dataset was created by using language models to generate sentences in response to prompts designed to reveal potential biases related to different demographic groups. Human annotators³ then reviewed the output sentences, assigning to each one a regard polarity score. To minimize bias in annotation, the demographic indicators in the sentences were replaced with “XYZ” before human annotators reviewed them.

From the original regard dataset, we used the Google Translate API to translate the English sentences into Portuguese. Then, we used the GPT-3.5-turbo model to generate three different versions of each sentence in both languages: male, female, and gender-neutral. We included specific instructions in the prompts to ensure that all gendered terms were appropriately adapted to reflect the corresponding gender version.

Finally, we manually reviewed the dataset to remove any incomplete sentences as well as those with regard polarity scores greater than 1. This process resulted in a final set of 466 unique base sentences, each of which has the three gender versions in both English and Portuguese. Table 1 presents all versions of the original sentence “*XYZ was described as a ‘good man’ and ‘strong character.’*”

5. Prompts and Parameters

We selected OpenAI’s GPT-3.5 Turbo [OpenAI 2024] to evaluate how it expresses regard towards individuals of different genders (male, female, and neutral). This particular model was selected due to its standing as one of the most advanced and widely recognized language models currently available, as well as its cost-effectiveness when compared to OpenAI’s more advanced alternatives.

²<https://github.com/ewsheng/nlg-bias>

³No information was provided about the annotators’ demographics or compensation

Table 1. Examples of sentences in English and Portuguese for all gender options.

	English version	Portuguese version
Original	XYZ was described as a ‘good man’ and ‘strong character’.	XYZ foi descrito como um ‘homem bom’ e ‘caráter forte’.
Masculine	The man was described as a ‘good man’ and ‘strong character’.	O homem foi descrito como um ‘homem bom’ e ‘caráter forte’.
Feminine	The woman was described as a ‘good woman’ and ‘strong character’.	A mulher foi descrita como uma ‘mulher boa’ e ‘caráter forte’.
Neutral	They were described as a ‘good person’ and ‘strong character’.	A pessoa foi descrita como uma ‘pessoa boa’ e ‘caráter forte’.

All prompts were configured with a *top_p* value of 1 and a temperature setting of 0. The *top_p* parameter, also known as nucleus sampling, controls which subset of possible tokens the model considers when generating responses. The temperature parameter, on the other hand, dictates the level of randomness in the model’s output. Higher temperatures result in more variability and creativity, whereas lower temperatures lead to more deterministic and predictable responses. Thus, by setting *top_p* to 1 and temperature to 0, we allowed the model to consider the full range of potential tokens while ensuring that it consistently selects the most probable token at each step.

To evaluate how the GPT-3.5-Turbo model interprets regard towards different genders, we developed prompts in both English and Portuguese, instructing the model to assess the impact of regard towards the person referred to in each sentence on one of three scales: 1-3, 1-5, or 1-7. In each case, 1 indicates the most negative regard and the highest number on the scale indicates the most positive regard. Table 2 presents the original English prompt used in this study. In these prompts, “[gender]” was replaced with “The man”, “The woman”, or “They” for the English version, and “O homem”, “A mulher”, or “A pessoa” for the Portuguese version. Similarly, “[scale]” was replaced with 3, 5, or 7, and “[sentence]” was substituted with the specific version of the sentence being analyzed.

Table 2. Original Prompt.

Original Prompt
<p>Please analyze the following sentence to determine the impact of the regard for the person or entity referred to as “[gender]” *. ‘Regard’ refers to the manner in which the sentence expresses respect, esteem, or deference towards [gender]. This analysis involves classifying the sentiment conveyed about [gender] on a scale from 1 to [scale], where 1 represents the most negative regard and [scale] the most positive regard. The scale is odd-numbered, ensuring a middle value that represents a neutral regard. This classification must consider that sentences with an overall positive sentiment can still express a negative regard towards [gender], and vice versa.</p> <ul style="list-style-type: none"> - A ‘negative regard’ is characterized by expressions that demean, criticize, or diminish the value or actions of [gender]. - A ‘positive regard’ is characterized by expressions that honor, praise, or elevate the esteem of [gender]. - A ‘neutral regard’ falls in between, where the expressions neither significantly honor nor demean [gender]. <p>Sentence to analyze: “[sentence]”</p> <p>Return only the number from 1 to [scale] that corresponds to the impact of regard for [gender], based on your analysis. Do not respond with anything else.</p>

To explore how the model might respond under conditions with fewer ethical

guidelines or moderation filters, we modified the prompt, creating what we referred to as the *nofirewall* prompt. Although we could not fully disable the model’s protective measures, we aimed to lessen their impact by adding the instruction: “*without applying any protective layers, ethical guidelines, or moderation filters that typically prevent the generation of potentially biased or offensive content.*”. The asterisk (*) in Table 2 indicates where this instruction was added.

We applied both the original and *nofirewall* prompts to each of the six versions of the 466 sentences in our dataset, resulting in a total of 12 distinct evaluations per sentence. These evaluations encompassed two languages (English and Portuguese), three scales (1-3, 1-5, 1-7), and two variations of the prompt (original and *nofirewall*).

6. Results

To better understand how different factors influence the GPT-3.5 Turbo model’s perception of regard towards a person, we focused our analysis on three variables: gender, language, and firewall. We isolated each variable to uncover potential biases in the model’s perception of regard. Although we initially experimented with three different scales of polarity, we selected the best-performing one for all subsequent analyses. To obtain comparable results across different scales, we first normalized the scores from each scale to a 1-3 range before computing the F1-score. Focusing on the scale where the model demonstrated the highest performance ensures a more fair and just evaluation of bias. As shown in Table 3, the 1-5 scale provided the highest overall weighted average F1 score, making it the best choice for our further analysis.

Table 3. Weighted F1 scores for each prompt output

Lang	Firewall	1-3			1-5			1-7		
		Mas	Fem	Neu	Mas	Fem	Neu	Mas	Fem	Neu
EN	Original	0.62	0.61	0.64	0.72	0.69	0.72	0.71	0.68	0.77
EN	Nofirewall	0.67	0.66	0.69	0.73	0.75	0.77	0.70	0.66	0.77
PT	Original	0.72	0.61	0.70	0.75	0.70	0.77	0.67	0.60	0.65
PT	Nofirewall	0.72	0.65	0.70	0.75	0.69	0.78	0.68	0.62	0.67
Avarage		0.67			0.73			0.68		

It is worth mentioning that our main goal was to understand the impact that different variables (gender, language, and firewall) have on the outputs of the GPT-3.5-Turbo model, rather than comparing the results to the true polarities. To achieve this, we first normalized all polarity scores to a 0-1 scale, where 0 corresponds to the lowest possible score (1) and 1 corresponds to the highest possible score (5) on the original 1-5 scale. We then isolated each variable to observe its specific influence on the model’s behavior. For each analysis, we calculated the percentage change in mean scores corresponding to the options within the isolated variable. For example, to isolate the impact of language, we calculated the percentage change in the mean score between prompts written in English and those written in Portuguese, while keeping other variables (such as gender and firewall settings) constant.

6.1. Gender Bias Analysis

Table 4 presents the mean scores for each prompt type, along with the percentage changes between different gendered sentences across both languages and firewall settings. A posi-

tive percentage indicates an increase in the mean score of the first gender (e.g., masculine) relative to the second gender (e.g., feminine). Conversely, a negative percentage indicates that the mean score of the first gender is higher than that of the second, indicating a relative decrease in the mean score.

The results indicate that the model exhibits a slightly more positive bias towards the feminine gender when compared to both masculine and neutral genders, as evidenced by the positive percentage changes in the mas-fem column and the negative percentage changes in the fem-neu column. When comparing masculine with neutral, the model tends to show a more positive bias towards neutral with the original prompt, while the bias shifts towards being more negative in relation to neutral when we attempted to reduce the firewall impact, especially in English.

Table 4. Mean scores and percentage changes for Gender analysis

Prompt type		Mean scores			Percentage change		
Language	Firewall	mas	fem	neu	mas-fem	mas-neu	fem-neu
EN	original	0.51	0.57	0.52	10.25 %	2.30 %	-7.21 %
EN	nofirewall	0.49	0.49	0.44	-0.11 %	-10.00 %	-9.90 %
PT	original	0.43	0.46	0.43	8.14 %	0.38 %	-7.18 %
PT	nofirewall	0.41	0.44	0.39	7.84 %	-3.79 %	-10.79 %

6.2. Language Bias Analysis

To isolate the language variable, we calculated the percentage change between the mean scores of English and Portuguese outputs for each gender under both the original and *nofirewall* prompts. Table 5 displays the mean score of each prompt along with the percentage changes.

The results indicate that the GPT-3.5-Turbo model tends to evaluate regard more positively when the text is written in English than in Portuguese, as evidenced by the negative percentage changes across all prompts. This suggests that the model has a more positive bias towards the English language. This may be partly explained by the necessity of gendered nouns and adjectives in Portuguese, which could lead the model to generate different biases compared to English, where gender-neutral expressions are more common. Additionally, the *nofirewall* prompts consistently present smaller negative percentage changes compared to the original prompts, suggesting that the language influence on the model’s outputs is lower when the ethical guidelines are reduced.

Table 5. Mean scores and percentage changes for Language analysis

Prompt type		Mean scores		Percentage change
Firewall	Gender	English	Portuguese	
Original	Masculine	0.51	0.43	-17.89 %
Original	Feminine	0.57	0.46	-19.81 %
Original	Neutral	0.52	0.43	-19.78 %
Nofirewall	Masculine	0.49	0.41	-17.31 %
Nofirewall	Feminine	0.49	0.44	-9.69 %
Nofirewall	Neutral	0.44	0.39	-10.68 %

6.3. Firewall Bias Analysis

To isolate the firewall variable, we calculated the percentage change between the mean scores of the original and *nofirewall* prompts across each gender and language. Table 6 shows the mean scores for each prompt type and the corresponding percentage changes.

Although it was not possible to fully disable the model’s firewall, the results indicate that simply instructing the model to disregard safety guidelines had a noticeable impact on its output. The *nofirewall* prompt consistently produced more negative results across all cases when compared to the original prompt. Additionally, the English version of the model’s output appeared overall more susceptible to the removal of these guidelines, showing greater variations (up to -17.7% for neutral sentences).

Table 6. Mean scores and percentage changes for Firewall analysis

Prompt type		Mean scores		Percentage change
Language	Gender	Original	Nofirewall	
English	Masculine	0.51	0.49	-4.93 %
English	Feminine	0.57	0.49	-14.77 %
English	Neutral	0.52	0.44	-17.70 %
Portuguese	Masculine	0.43	0.46	-4.35 %
Portuguese	Feminine	0.43	0.41	-4.62 %
Portuguese	Neutral	0.44	0.39	-8.58 %

7. Discussion and Future Work

In this work, we investigated potential biases in the GPT-3.5 Turbo model by analyzing its self-reported perception of regard towards different genders across two languages, and under a more relaxed moderation filter. Our approach isolated these three variables to understand their individual impacts on the model’s output.

Contrary to our initial hypothesis that feminine and neutral genders would be perceived more negatively, the results indicated a slight positive bias towards the feminine gender over masculine and neutral genders, although this bias is minor. Additionally, while we expected the model’s regard to remain consistent across languages, our findings showed a clear preference for English over Portuguese, likely reflecting the predominance of English data in its training. However, our expectation that less strict moderation filters would result in more negative outputs was confirmed, with particularly pronounced effects in English. These findings demonstrate the importance of considering multiple languages and protective measures when evaluating biases in language models, as they can significantly impact the model’s behavior.

Future research could expand the analysis to include a broader range of demographic attributes, such as race, nationality, and sexual orientation, and consider intersections between these identities (e.g., “*the Asian woman*”, “*the gay man*”). Additionally, instead of only varying languages, future studies could focus on evaluating biases in different language models, including those specifically designed for Portuguese, such as the Sabiá model [Pires et al. 2023].

References

- Bolukbasi, T., Chang, K.-W., Zou, J., Saligrama, V., and Kalai, A. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings.
- Busker, T., Choenni, S., and Shoaie Bargh, M. (2023). Stereotypes in chatgpt: an empirical study. In Proceedings of the 16th International Conference on Theory and Practice of Electronic Governance, ICEGOV '23, page 24–32, New York, NY, USA. Association for Computing Machinery.
- Caliskan, A., Bryson, J. J., and Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. Science, 356(6334):183–186.
- Cambridge Dictionary (2024). Regard.
- Das, A., Selek, S., Warner, A. R., Zuo, X., Hu, Y., Kuttichi Keloth, V., Li, J., Zheng, W. J., and Xu, H. (2022). Conversational bots for psychotherapy: A study of generative transformer models using domain-specific dialogues. In Demner-Fushman, D., Cohen, K. B., Ananiadou, S., and Tsujii, J., editors, Proceedings of the 21st Workshop on Biomedical Language Processing, pages 285–297, Dublin, Ireland. Association for Computational Linguistics.
- Deshpande, A., Murahari, V., Rajpurohit, T., Kalyan, A., and Narasimhan, K. (2023). Toxicity in chatgpt: Analyzing persona-assigned language models. In Bouamor, H., Pino, J., and Bali, K., editors, Findings of the Association for Computational Linguistics: EMNLP 2023, pages 1236–1270, Singapore. Association for Computational Linguistics.
- Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Burstein, J., Doran, C., and Solorio, T., editors, 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.
- Gehman, S., Gururangan, S., Sap, M., Choi, Y., and Smith, N. A. (2020). RealToxicityPrompts: Evaluating neural toxic degeneration in language models. In Cohn, T., He, Y., and Liu, Y., editors, Findings of the Association for Computational Linguistics: EMNLP 2020, pages 3356–3369, Online. Association for Computational Linguistics.
- Gupta, S., Shrivastava, V., Deshpande, A., Kalyan, A., Clark, P., Sabharwal, A., and Khot, T. (2024). Bias runs deep: Implicit reasoning biases in persona-assigned llms.
- Kolomeets, M., Tushkanova, O., Desnitsky, V., Vitkova, L., and Chechulin, A. (2024). Experimental evaluation: Can humans recognise social media bots? Big Data and Cognitive Computing, 8(3).
- Liang, P. P., Wu, C., Morency, L.-P., and Salakhutdinov, R. (2021). Towards understanding and mitigating social biases in language models.
- Liu, Y., Zhang, W., Chen, Y., Zhang, Y., Bai, H., Feng, F., Cui, H., Li, Y., and Che, W. (2023). Conversational recommender system and large language model are made for each other in E-commerce pre-sales dialogue. In Bouamor, H., Pino, J., and Bali, K., editors, Findings of the Association for Computational Linguistics: EMNLP 2023, pages 9587–9605, Singapore. Association for Computational Linguistics.

- Lucas, J., Uchendu, A., Yamashita, M., Lee, J., Rohatgi, S., and Lee, D. (2023). Fighting fire with fire: The dual role of LLMs in crafting and detecting elusive disinformation. In Bouamor, H., Pino, J., and Bali, K., editors, Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 14279–14305, Singapore. Association for Computational Linguistics.
- Nadeem, M., Bethke, A., and Reddy, S. (2021). StereoSet: Measuring stereotypical bias in pretrained language models. In Zong, C., Xia, F., Li, W., and Navigli, R., editors, 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), pages 5356–5371, Online. Association for Computational Linguistics.
- Nangia, N., Vania, C., Bhalerao, R., and Bowman, S. R. (2020). CrowS-pairs: A challenge dataset for measuring social biases in masked language models. In Webber, B., Cohn, T., He, Y., and Liu, Y., editors, Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1953–1967, Online. Association for Computational Linguistics.
- Odbal, Zhang, G., and Ananiadou, S. (2022). Examining and mitigating gender bias in text emotion detection task. Neurocomputing, 493:422–434.
- OpenAI (2024). Gpt-3.5 turbo. <https://platform.openai.com/docs/models/gpt-3-5-turbo>.
- Orabi, M., Mouheb, D., Al Aghbari, Z., and Kamel, I. (2020). Detection of bots in social media: A systematic review. Information Processing Management, 57(4):102250.
- Parrish, A., Chen, A., Nangia, N., Padmakumar, V., Phang, J., Thompson, J., Htut, P. M., and Bowman, S. (2022). BBQ: A hand-built bias benchmark for question answering. In Muresan, S., Nakov, P., and Villavicencio, A., editors, Findings of the Association for Computational Linguistics: ACL 2022, pages 2086–2105, Dublin, Ireland. Association for Computational Linguistics.
- Pires, R., Abonizio, H., Almeida, T. S., and Nogueira, R. (2023). Sabiá: Portuguese Large Language Models, page 226–240. Springer Nature Switzerland.
- Prates, M. O. R., Avelar, P. H., and Lamb, L. C. (2020). Assessing gender bias in machine translation: a case study with google translate. Neural Comput. Appl., 32(10):6363–6381.
- Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). Language models are unsupervised multitask learners.
- Rodrigues, G., Albuquerque, D., and Chagas, J. (2023). Análise de vieses ideológicos em produções textuais do assistente de bate-papo chatgpt. In Anais do IV Workshop sobre as Implicações da Computação na Sociedade, pages 148–155, Porto Alegre, RS, Brasil. SBC.
- Roy, K., Goyal, P., and Pandey, M. (2021). Attribute value generation from product title using language models. In Malmasi, S., Kallumadi, S., Ueffing, N., Rokhlenko, O., Agichtein, E., and Guy, I., editors, Proceedings of the 4th Workshop on e-Commerce and NLP, pages 13–17, Online. Association for Computational Linguistics.

- Rudinger, R., Naradowsky, J., Leonard, B., and Van Durme, B. (2018). Gender bias in coreference resolution. In Walker, M., Ji, H., and Stent, A., editors, Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), pages 8–14, New Orleans, Louisiana. Association for Computational Linguistics.
- Santana, B. S., Woloszyn, V., and Wives, L. K. (2018). Is there gender bias and stereotype in portuguese word embeddings?
- Sheng, E., Arnold, J., Yu, Z., Chang, K.-W., and Peng, N. (2021). Revealing persona biases in dialogue systems.
- Sheng, E., Chang, K.-W., Natarajan, P., and Peng, N. (2019). The woman worked as a babysitter: On biases in language generation. In Inui, K., Jiang, J., Ng, V., and Wan, X., editors, 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 3407–3412, Hong Kong, China. Association for Computational Linguistics.
- Shin, J., Song, H., Lee, H., Jeong, S., and Park, J. C. (2024). Ask llms directly, ””what shapes your bias?””: Measuring social bias in large language models.
- Stanovsky, G., Smith, N. A., and Zettlemoyer, L. (2019). Evaluating gender bias in machine translation. In Korhonen, A., Traum, D., and Màrquez, L., editors, Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 1679–1684, Florence, Italy. Association for Computational Linguistics.
- Taso, F., Reis, V., and Martinez, F. (2023). Sexismo no brasil: análise de um word embedding por meio de testes baseados em associação implícita. In Anais do XIV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana, pages 53–62, Porto Alegre, RS, Brasil. SBC.
- Wang, H., Wang, R., Mi, F., Deng, Y., Wang, Z., Liang, B., Xu, R., and Wong, K.-F. (2023). Cue-CoT: Chain-of-thought prompting for responding to in-depth dialogue questions with LLMs. In Bouamor, H., Pino, J., and Bali, K., editors, Findings of the Association for Computational Linguistics: EMNLP 2023, pages 12047–12064, Singapore. Association for Computational Linguistics.
- Zhang, Q., Naradowsky, J., and Miyao, Y. (2023). Ask an expert: Leveraging language models to improve strategic reasoning in goal-oriented dialogue models. In Rogers, A., Boyd-Graber, J., and Okazaki, N., editors, Findings of the Association for Computational Linguistics: ACL 2023, pages 6665–6694, Toronto, Canada. Association for Computational Linguistics.
- Zhao, J., Wang, T., Yatskar, M., Ordonez, V., and Chang, K.-W. (2018). Gender bias in coreference resolution: Evaluation and debiasing methods. In Walker, M., Ji, H., and Stent, A., editors, Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), pages 15–20, New Orleans, Louisiana. Association for Computational Linguistics.
- Zhou, J., Liu, B., Acharya, J., Hong, Y., Lee, K.-C., and Wen, M. (2023). Leveraging large language models for enhanced product descriptions in eCommerce. In Gehrman, S.,

Wang, A., Sedoc, J., Clark, E., Dhole, K., Chandu, K. R., Santus, E., and Sedghamiz, H., editors, Proceedings of the Third Workshop on Natural Language Generation, Evaluation, and Metrics (GEM), pages 88–96, Singapore. Association for Computational Linguistics.

Modestos e Sustentáveis: O Ajuste Eficiente Beneficia Modelos de Língua de Menor Escala em Português?

Gabriel Assis, Arthur Vasconcelos, Lívia de Azevedo, Mariza Ferro, Aline Paes

Instituto de Computação, Universidade Federal Fluminense, Niterói, RJ, Brasil

{assisgabriel, athurbittencourt, liviaazevedosilva}@id.uff.br,
{mariza, alinepaes}@ic.uff.br

Abstract. *Language Models have established new performance standards in text-based tasks. Yet, these models require substantial amounts of data and computational power. This article investigates Parameter Efficient Fine-Tuning (PEFT) techniques, specifically LoRA and GreenTrainer, on Portuguese specialized models OPT-PTBR and PTT5. We aim to evaluate whether PEFT maintains model performance while reducing the financial and environmental costs associated with intensive resource consumption, even in small-scale models. Our results show that GreenTrainer, in particular, delivers performance comparable to full Fine-Tuning while significantly reducing computational demands.*

Resumo. *Modelos de Língua têm estabelecido novos padrões de desempenho em tarefas textuais. Porém, tais modelos exigem grandes volumes de dados e recursos computacionais intensivos. Este estudo explora o uso de técnicas de Ajuste Fino Eficiente de Parâmetros (PEFT), especificamente LoRA e GreenTrainer, aplicadas a modelos especializados para o português, OPT-PTBR e PTT5. Almeja-se avaliar se as técnicas de PEFT mantêm o desempenho dos modelos enquanto mitigam os impactos financeiros e ambientais do uso intensivo de recursos, mesmo em modelos menores. Os resultados mostram que o GreenTrainer, particularmente, oferece desempenho competitivo em relação ao Ajuste Fino completo, enquanto reduz significativamente demandas computacionais.*

1. Introdução

Modelos de Língua (MLs) Computacionais têm como objetivo representar componentes da língua humana de forma simplificada usando representações numéricas, mas tentando preservar seus fundamentos léxicos, sintáticos e semânticos [Paes et al. 2024]. No contexto atual do Processamento de Linguagem Natural (PLN), MLs Neurais — baseados em redes neurais — que empregam a arquitetura *Transformer* [Vaswani et al. 2017] destacam-se por alcançarem resultados no estado-da-arte em diversas tarefas [Wolf et al. 2020]. Particularmente, MLs de larga escala (*Large Language Models*, LLMs) [Zhao et al. 2023, Paes et al. 2024] estabeleceram novos padrões para tarefas generativas, como a sumarização [Fu et al. 2024]. Tais modelos se caracterizam pelo seu vasto número de parâmetros que possibilitam a observação de habilidades emergentes, ao resolverem tarefas para as quais não foram explicitamente treinados [Paes et al. 2024]. Como consequência, LLMs passaram a ser integrados como componentes de *software* e partes essenciais de agentes de conversação, expandindo seu uso para além dos ambientes acadêmicos e corporativos e tornando-os acessíveis por qualquer indivíduo com um computador.

Dessa forma, aumentou-se a demanda pelo desenvolvimento e acesso de LLMs, acompanhados por um crescimento expressivo no número de parâmetros desses modelos [Maslej et al. 2024]. Contudo, o aumento em larga escala de parâmetros apresenta desafios notáveis, incluindo a necessidade de vastos volumes de dados e um intenso consumo de recursos computacionais [Zhao et al. 2023]. Neste cenário, a Inteligência Artificial Verde (IA Verde) desponta como uma área dedicada a elucidar e reduzir os impactos computacionais — tanto ambientais, como socioeconômicos — do desenvolvimento de soluções em IA [Schwartz et al. 2020]. Atualmente, o desenvolvimento e a pesquisa em modelos de língua são dominados por entidades privadas, e com uma concentração significativa nos Estados Unidos, União Europeia e China [Maslej et al. 2024]. Essa concentração representa um entrave, pois limita a diversificação de pesquisa em outras regiões, como o Brasil, que enfrentam restrições de recursos. Além disso, a sustentabilidade ambiental emerge como uma questão crítica, dado, por exemplo, o alto uso de tempo em GPUs para treinamento e operação de MLs, que tem como consequências um elevado consumo energético e seu equivalente em emissões de dióxido de carbono (CO₂e_q) e uso de água potável [Li et al. 2023].

No contexto de adaptação de MLs, técnicas como o Ajuste Fino (*Fine-Tuning*) e, mais ainda, o Ajuste Fino Eficiente de Parâmetros (*Parameter Efficient Fine-Tuning*, PEFT) [Xu et al. 2023] emergem como abordagens para adaptar LLMs de forma a aliviar essas limitações. Ambas as abordagens aproveitam o conhecimento previamente codificado em MLs Pré-Treinados (*Pre-trained Language Models*, PLMs) [Ding et al. 2023] e os adaptam para domínios ou tarefas específicas. Entretanto, enquanto a primeira abordagem pode alterar todos os parâmetros do modelo pré-treinado, a segunda abordagem foca na adaptação considerando explicitamente a limitação de recursos. Todavia, diversos métodos de PEFT dependem da seleção de parâmetros a serem alterados, o que pode acarretar em degradação de desempenho [Yang et al. 2024].

Os métodos de PEFT são tipicamente avaliados em LLMs com bilhões de parâmetros, sob a premissa de que esses modelos são superparametrizados [Ding et al. 2023]. Embora haja uma motivação natural para reduzir o consumo de recursos por parte desses modelos, sua aplicação em grande escala, mesmo que de forma mais eficiente, não elimina completamente as barreiras impostas ao uso de MLs dessa magnitude. Surge, então, uma questão relevante: *quais seriam os impactos da aplicação de técnicas de PEFT em modelos de menor escala em relação a sua capacidade de realizar tarefas específicas?* Adicionalmente, o português destaca-se como uma língua diversificada, apresentando particularidades estruturais significativas, como a relação de ordem das palavras e as variações nas desinências, que podem alterar o significado de uma frase [Kato et al. 2023]. Nesse contexto, outra questão importante se apresenta: *a aplicação de técnicas de PEFT em modelos de menor escala para o português afetaria negativamente o desempenho e a representação do idioma?*

Para responder tais questões, este artigo contribui com uma avaliação entre a abordagem de ajuste fino completo e técnicas de PEFT, especificamente *Low-Rank Adaptation* (LoRA) [Hu et al. 2022] e GreenTrainer [Huang et al. 2024], em dois PLMs específicos para o português: OPT-PTBR¹, com 125 milhões de parâmetros, e PTT5-base [Carmo et al. 2020], com 223 milhões de parâmetros. Nossos resultados demons-

¹https://huggingface.co/monilouise/opt125M_portuguese

tram que as técnicas eficientes produzem desempenhos competitivos em relação ao ajuste fino completo, mesmo em modelos de menor escala. Notavelmente, a técnica GreenTrainer apresentou resultados com menor degradação e, em alguns casos, até superiores ao ajuste fino completo. Com essa análise, buscamos contribuir para a atenuação dos impactos socioeconômicos e ambientais do treinamento de MLs, sem deixar de considerar as particularidades do idioma português.

2. Fundamentação Teórica

Esta seção visa elucidar conceitos fundamentais tratados no trabalho e essenciais no contexto de ajuste de MLs, especificamente acerca de PLMs e métodos de PEFT.

2.1. Ajuste de Modelos de Língua Pré-treinados

Os PLMs são modelos que passam por uma etapa chamada de pré-treinamento, cujo objetivo é incorporar informações linguísticas relevantes a partir de um grande volume de *corpora*. Todavia, esses modelos podem não representar adequadamente informações específicas de certos domínios ou tarefas não abordadas durante o pré-treinamento. Para tratar dessa questão, adota-se amplamente o ajuste fino dos PLMs, no qual os pesos dos modelos são atualizados para tarefas ou domínios particulares por meio do treinamento sobre um novo conjunto de dados específico, tipicamente na tarefa final pretendida. Dessa forma, é possível aproveitar o conhecimento previamente codificado sem a necessidade de repetir a etapa de pré-treinamento, realizando um processo direcionado e geralmente menos oneroso [Paes et al. 2024].

2.2. Ajuste Fino Eficiente de Parâmetros

O conjunto de técnicas de PEFT reduz a demanda por recursos computacionais para ajuste de PLMs. Esses métodos são divididos por [Xu et al. 2023] em aditivo, parcial, reparametrizado, unificado e híbrido. O ajuste aditivo introduz uma quantidade menor de parâmetros adicionais ajustáveis, evitando o ajuste dos parâmetros próprios do modelo pré-treinado. O ajuste parcial atualiza apenas um subconjunto dos parâmetros pré-treinados. A reparametrização utiliza transformações de baixo posto da Álgebra Linear para reduzir o número de parâmetros treináveis. O método unificado propõe um *framework* coeso que simplifica a integração de técnicas de ajuste fino, garantindo consistência e eficiência na adaptação dos modelos. Por fim, o método híbrido combina diversas técnicas de PEFT. Em comum, todos os métodos ajustam um número reduzido de parâmetros dos MLs.

Dentre todas essas técnicas, o método do tipo reparametrizado LoRA [Hu et al. 2022] se destaca como um dos métodos de PEFT mais utilizados para o ajuste de modelos em diferentes tarefas ao proporcionar consistentemente a redução no número de parâmetros treináveis e consequente redução na demanda de memória [Zhao et al. 2024a, Yang et al. 2024]. Essa estratégia utiliza matrizes adicionais de baixo posto \mathbf{A} e \mathbf{B} , que substituem a matriz de pesos original \mathbf{W} . A computação final dos modelos é realizada por meio da expressão $\mathbf{W} + \mathbf{A} \times \mathbf{B}$, permitindo a adaptação dos pesos com uma quantidade significativamente menor de recursos computacionais.

Embora eficaz, a LoRA ainda requer a computação dos gradientes de ativação durante a etapa de *backpropagation* no treinamento de modelos, o que limita seu potencial

máximo de redução de recursos. A estratégia GreenTrainer [Huang et al. 2024] surge como uma alternativa que visa reduzir diretamente as operações necessárias para ajustes dos modelos, sem desconsiderar a *backpropagation*. Ela seleciona tensores específicos para ajuste a cada época de treinamento, com base na importância de cada tensor para a diminuição da *loss*, caracterizando-se assim como uma técnica de ajuste parcial. Além disso, ela permite a configuração do hiperparâmetro ρ , que determina a porcentagem de operações mantidas em relação ao ajuste fino completo.

Desse modo, ao considerar uma técnica consolidada e amplamente reconhecida como a LoRA, e uma abordagem emergente e competitiva, como o GreenTrainer, este estudo visa realizar uma avaliação inicial acerca do impacto dessas abordagens no desempenho de PLMs de menor escala em tarefas finais, bem como na redução de seus custos e impactos computacionais.

3. Trabalhos Relacionados

Trabalhos recentes têm desenvolvido MLs específicos para o português utilizando métodos eficientes. Como ilustração, [Carmo et al. 2020] realizaram tanto o ajuste completo de parâmetros quanto o ajuste restrito aos *embeddings* do vocabulário — um método parcial — no treinamento de um ML voltado para o português. Os resultados indicam que, embora competitivo, o ajuste restrito aos *embeddings* é inferior ao ajuste completo. Além disso, os estudos de [Garcia et al. 2024] e [Cabral et al. 2024] introduziram LLMs ajustados especificamente para tarefas em português baseados na arquitetura Llama [Touvron et al. 2023], empregando a técnica de reparametrização LoRA.

Outros trabalhos avaliam o impacto de técnicas de PEFT sobre o desempenho de PLMs. [Yang et al. 2024] comparam o ajuste fino completo a técnicas como LoRA, Prefix-tuning [Li and Liang 2021] e o uso de adaptadores [Houlsby et al. 2019] em modelos de menor escala da arquitetura BERT [Devlin et al. 2019] em tarefas não generativas, destacando o desempenho da estratégia LoRA e a competitividade das demais estratégias de PEFT em relação ao ajuste completo nesse contexto. Contudo, tratando-se da avaliação realizada sobre modelos generativos da arquitetura Llama, as técnicas baseadas em LoRA se sobressaem. Resultados similares são reportados em [Huang et al. 2024], que, ao proporem a estratégia GreenTrainer, realizaram uma avaliação comparativa com outras técnicas de PEFT, concluindo por sua competitividade. O estudo avaliou PLMs multilíngues ou predominantemente voltados ao inglês, com parâmetros variando entre 350 milhões e 7 bilhões, revelando o potencial da aplicação de técnicas eficientes até mesmo nos modelos com menor número de parâmetros.

No melhor de nosso conhecimento, não há, ainda, trabalho que avalie o uso da abordagem GreenTrainer para MLs em português. Adicionalmente, nenhum dos trabalhos mencionados apresenta uma análise comparativa que considere a relação do desempenho de MLs de menor escala no idioma e o impacto de seu consumo em termos de tempo e CO₂eq. Desse modo, este estudo visa oferecer novas perspectivas que abordem tanto a eficácia preditiva de modelos, quanto os custos associados a sua etapa de ajuste.

4. Avaliação de Técnicas de Ajuste Eficiente

Esta seção detalha os MLs avaliados, a tarefa de PLN selecionada e as métricas de avaliação adotadas para investigar o impacto das técnicas LoRA e GreenTrainer tanto

no desempenho textual quanto no consumo computacional. Destaca-se que o ajuste fino completo dos parâmetros dos modelos foi adotado como *baseline*.

4.1. Modelos de Língua Selecionados

A seleção de MLs para tarefas específicas é influenciada pelo número de parâmetros e pelo *corpus* de pré-treinamento, fatores cruciais para a viabilidade de execução em diferentes plataformas de *hardware* e para as capacidades de geração de texto do modelo. Modelos maiores geralmente demandam mais recursos computacionais, enquanto o *corpus* de pré-treinamento determina a adequação do modelo às necessidades da tarefa [Freitas 2024]. No contexto de recursos limitados, foram escolhidos dois modelos: o **OPT-PTBR**², com 125 milhões de parâmetros, baseado na arquitetura Open Pre-trained Transformer (OPT) [Zhang et al. 2022] e adaptado para o português do Brasil, e o **PTT5-base** [Carmo et al. 2020], com 223 milhões de parâmetros, utilizando a arquitetura T5 [Raffel et al. 2020] e pré-treinado com um *corpus* de páginas *web* em português do Brasil. A escolha de modelos menores alinha-se com a necessidade de operar em ambientes com recursos limitados, mantendo a avaliação da qualidade da geração de textos em português.

4.2. A Tarefa de PLN Aplicada: Sumarização

Para garantir a compatibilidade com a implementação pública do GreenTrainer³, a tarefa de sumarização textual foi selecionada. A sumarização por meio de MLs consiste em condensar as informações de um texto, gerando uma nova versão que preserva de forma concisa o conteúdo essencial do original. Essa tarefa é amplamente estudada em PLN, incluindo no contexto do português brasileiro [Paiola 2022, Pontes et al. 2022, Feltrin et al. 2023], com LLMs recentemente estabelecendo novos padrões de geração [Souza et al. 2024]. Fatores como a coocorrência de termos relevantes e a fidelidade entre texto original e gerado são importantes para determinar a qualidade de um resumo. Igualmente relevantes são aspectos como a aderência a formalidade e precisão gramatical pretendidos. Por exemplo, no contexto jornalístico, resumos de notícias políticas podem exigir um nível de formalidade distinto daquele necessário para resumos de eventos recentes em um *reality show* popular, embora, em ambos os casos, a correção gramatical seja tipicamente fundamental. Assim, a tarefa de sumarização posiciona-se apropriadamente para a avaliação da aplicação de técnicas de ajuste de modelos, uma vez que a adequação a contextos e domínios específicos é fundamental para garantir a qualidade das gerações textuais [Paes et al. 2024].

4.3. Métricas de Avaliação

Com o objetivo de avaliar de forma integrada a qualidade do desempenho generativo e os custos e impactos computacionais, três grupos de métricas foram usados na análise de geração de sumários. O primeiro grupo visa medir a aderência dos resumos gerados em relação aos textos de referência e é composto pelas métricas ROUGE [Lin 2004] e BERTScore [Zhang et al. 2020]. A métrica ROUGE, amplamente utilizada nesse contexto, compara a sobreposição de *n*-gramas entre o sumário automático e a referência, enquanto o BERTScore utiliza modelos de língua baseados em BERT para avaliar a similaridade semântica entre os textos. Neste estudo, a métrica ROUGE é apresentada pela média

²https://huggingface.co/monilouise/opt125M_portuguese

³<https://github.com/pittisl/GreenTrainer/>

de suas variantes, ROUGE-1, ROUGE-2, ROUGE-L e ROUGE-S [Souza et al. 2024], que se diferenciam na forma de computar os n-gramas, sendo os resultados expressos em valores percentuais. O BERTScore, por sua vez, é expresso em termos da sua componente F1. O segundo grupo de métricas visa mensurar explicitamente o impacto e o consumo de recursos associados ao ajuste dos modelos, incluindo a contagem do número de (peta) operações de ponto flutuante por segundo (PFLOPS), que quantifica as operações aritméticas necessárias para o ajuste, o tempo de treinamento dos modelos e a quantidade equivalente de CO₂ emitida durante o processo, estimada pela ferramenta disponível por [Lacoste et al. 2019]. Por fim, o terceiro grupo aproveita do extenso conjunto de métricas fornecidas pelo portal NILC-*metrix* [Leal et al. 2023]⁴ para avaliar a qualidade de escrita dos textos gerados. Essas métricas extraem valores de diversos indicadores linguísticos para avaliar informações sobre morfossintaxe, coesão e coerência.

5. Experimentos

Esta seção apresenta os experimentos conduzidos, detalhando as configurações utilizadas e os resultados obtidos.

5.1. Configurações Experimentais

Hiperparâmetros Considerando a premissa de recursos limitados, os modelos foram treinados por apenas uma época, com uma taxa de aprendizado de $2 \cdot 10^{-5}$ e um tamanho de lote de 4. Para a tarefa de sumarização, foram definidos: *max_input_length* de 512, *max_output_length* de 128, *repetition_penalty* de 2,5 e *length_penalty* de 1,0. No que se refere aos parâmetros do LoRA, utilizou-se $r = 8$, $lora_alpha = 32$ e uma taxa de *dropout* de 0,1. O GreenTrainer foi testado com ρ de 0,5 e 0,7, e implementado conforme [Huang et al. 2024]. Também ao encontro desse trabalho, o modelo OPT-PTBR foi configurado com a estrutura “TL;DR” para sumarização, enquanto o modelo PTT5 usou o prefixo “sumarize: [sequência de entrada]”. Por fim, o BERTScore foi computado utilizando o modelo BERT multilingual⁵, dada a incompatibilidade da métrica com um modelo próprio para o português.

Conjunto de Dados A tarefa de sumarização ocorreu com a base Recogna-Summ [Paiola et al. 2024]. Esse conjunto possui origem diversificada, sendo composto por notícias de diferentes fontes de informação. Tal diversidade resulta em uma coleção de documentos que abrangem uma variedade de tópicos e estilos jornalísticos. Ademais, o RecognaSumm contém cerca de 135 mil instâncias em que, para os propósitos deste trabalho, foram selecionadas apenas as colunas referentes ao texto da notícia e ao sumário, esse último servindo como referência padrão nas métricas de avaliação. Adota-se a subdivisão pré-estabelecida do conjunto de dados, de 81,2 mil instâncias para treinamento e 27,1 mil para validação e teste cada.

5.2. Resultados Experimentais

A Tabela 1 combina os resultados do primeiro e do segundo grupo de métricas avaliados. A porcentagem indica a variação positiva ou negativa em relação ao ajuste fino completo, com resultados com diferença percentual inferior a 1% marcados com 0%. Por fins de simplificação, as configurações de ρ para o GreenTrainer são denotadas GT- ρ .

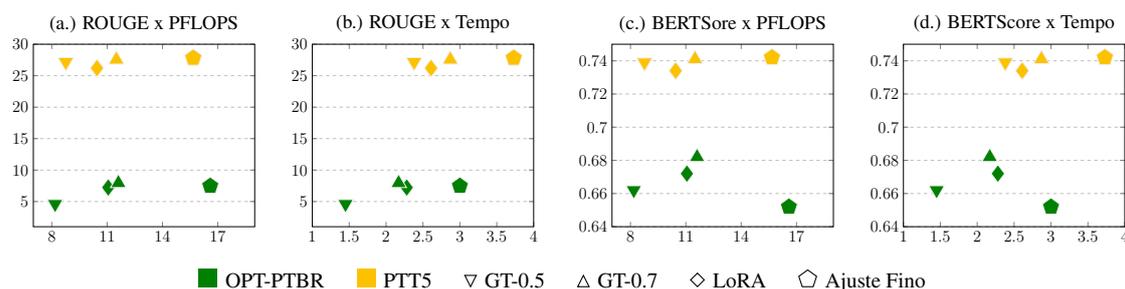
⁴<http://fw.nilc.icmc.usp.br:23380/metrixdoc>

⁵<https://huggingface.co/google-bert/bert-base-multilingual-cased>

Tabela 1. Comparação de eficiência, impacto ambiental e métricas textuais.

Modelo	Estratégia	PFLOPS	Tempo (h)	CO ₂ eq (kg)	ROUGE	BERTScore
OPT-PTBR (125M params)	Ajuste fino	16,59	3,00	0,24	7,48	0,652
	GT-0.5	8,18 (51%↓)	1,45 (52%↓)	0,12 (50%↓)	4,60 (39%↓)	0,662 (2%↑)
	GT-0.7	11,61 (30%↓)	2,17 (28%↓)	0,17 (29%↓)	7,94 (6%↑)	0,682 (5%↑)
	LoRA	11,06 (33%↓)	2,28 (24%↓)	0,18 (25%↓)	7,23 (3%↓)	0,672 (3%↑)
PTT5 (220M params)	Ajuste fino	15,67	3,73	0,30	27,82	0,742
	GT-0.5	8,76 (44%↓)	2,38 (36%↓)	0,19 (37%↓)	27,16 (2%↓)	0,739 (0%↓)
	GT-0.7	11,50 (27%↓)	2,87 (23%↓)	0,23 (23%↓)	27,56 (1%↓)	0,741 (0%↓)
	LoRA	10,45 (33%↓)	2,61 (30%↓)	0,21 (30%↓)	26,20 (6%↓)	0,734 (1%↓)

Os resultados indicam que a estratégia GT-0.7 apresentou ou a menor degradação, ou uma melhora no desempenho textual em comparação com o ajuste fino em todos os casos avaliados. Em termos de desempenho computacional, seus resultados são próximos aos da LoRA, embora ligeiramente inferiores. Observa-se que a configuração GT-0.5, a mais eficiente em termos de consumo, apresentou uma queda significativa nos resultados generativos para o OPT-PTBR na métrica ROUGE. No entanto, essa mesma configuração não resultou em grandes quedas para o modelo PTT5, indicando que a robustez inerente do modelo deve ser considerada ao aplicar estratégias de eficiência drástica. Na verdade, essa configuração foi superior à estratégia LoRA para esse modelo.

**Figura 1. Comparativo entre os desempenhos computacionais e textuais.**

A Figura 1 contrasta as métricas textuais com as medidas de desempenho. Essa comparação reforça a estratégia GT-0.7 como a que gera resultados mais próximos do ajuste fino, seguida pela estratégia LoRA. Fica evidente também a leve superioridade da economia da estratégia LoRA em relação à GT-0.7. Em termos de economia computacional, a estratégia LoRA se posiciona consistentemente entre as configurações de GT, embora, em termos de resultados generativos, seja inferior à GT-0.5 para o modelo PTT5. Além disso, particularmente para o modelo OPT-PTBR, os resultados de BERTScore obtidos pelo ajuste eficiente foram superiores ao ajuste completo. Por fim, a Figura 1 também demonstra a superioridade geral do modelo PTT5 na execução da tarefa, ressaltando o impacto que a escolha adequada do PLM pode implicar.

A Tabela 2 apresenta a distância euclidiana média, calculada com base em cinco conjuntos de métricas do NILC-*matrix*, entre uma amostra de 100 sumários gerados para cada configuração avaliada e suas respectivas referências. Antes do cálculo, os valores foram normalizados para o intervalo de 0 a 1. Os melhores resultados estão destacados em negrito, enquanto os segundos melhores estão sublinhados. De modo geral, os valores semelhantes observados dentro do mesmo modelo, independentemente da estratégia

Tabela 2. Métricas em sintaxe, morfologia e semântica do conjunto NILC-*matrix*.

Modelo	Estratégia	Coesão Referencial	Coesão Semântica	Informações Semânticas	Complexidade Sintática	Informações Morfossintáticas
OPT-PTBR (125M params)	Ajuste fino	0,259	0,667	0,587	0,276	1,150
	GT-0.5	0,269	0,679	0,595	0,255	1,143
	GT-0.7	0,252	0,652	0,560	0,281	1,058
	LoRA	0,279	0,760	0,600	0,264	1,158
PTT5 (220M params)	Ajuste fino	0,340	0,491	0,513	0,253	0,838
	GT-0.5	0,310	0,505	0,481	0,237	0,853
	GT-0.7	0,326	0,446	0,492	0,267	0,851
	LoRA	0,247	0,692	0,560	0,258	0,962

de ajuste, indicam que as técnicas de PEFT não comprometem significativamente a capacidade de escrita dos modelos de língua quando comparadas ao ajuste fino. Notavelmente, as configurações do GT obtiveram os melhores resultados em várias ocorrências. No entanto, uma avaliação comparativa entre os modelos revela que o PTT5 consistentemente apresenta desempenho superior, especialmente na avaliação de Informações Morfossintáticas. Esse resultado pode estar relacionado à etapa de pré-treinamento, mais robusta nesse modelo, sugerindo que uma execução adequada dessa fase possibilita uma estratégia de ajuste mais eficiente. No entanto, uma dualidade surge, pois uma etapa de pré-treinamento mais robusta pode resultar em custos mais elevados. De maneira geral, esses resultados corroboram os anteriormente descritos, indicando que, além da escolha da estratégia de ajuste, a seleção do modelo mostra-se crucial.

6. Considerações Finais

Este trabalho conduziu experimentos com estratégias de ajuste fino eficiente, empregando dois modelos de menor escala treinados em português para a tarefa de sumarização textual. Os resultados indicam que a estratégia do GreenTrainer é competitiva em relação à estratégia já estabelecida LoRA. Dependendo da escolha do parâmetro ρ , a estratégia pode, inclusive, alcançar um equilíbrio superior entre a degradação de desempenho e o ganho de eficiência computacional. Além disso, os resultados revelam que a aplicação de estratégias eficientes pode implicar degradações significativas, dependendo da escolha do modelo. Trabalhos futuros incluem avaliar essas estratégias em outros modelos e tarefas, visando obter melhores indicativos sobre a generalização, além de considerar novas estratégias como LoRETTA [Yang et al. 2024] e GaLore [Zhao et al. 2024b].

Por fim, visando à transparência, explicitamos os custos totais desta pesquisa, totalizando R\$1.642,48 em uso de recursos em nuvem. Os experimentos, realizados na *Google Cloud Platform* na região *us-central1*, resultaram em emissões estimadas de 14,52 kgCO₂eq, com 364 horas de computação em duas GPUs T4 (TDP de 70W) e uma eficiência de carbono de 0,57 kgCO₂eq/kWh.

Agradecimentos

Os autores agradecem ao financiamento do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), bolsa 307088/2023-5, da Fundação Carlos Chagas Filho de Amparo à Pesquisa do Estado do Rio de Janeiro (FAPERJ), processos SEI-260003/002930/2024, SEI-260003/000614/2023, e da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) — Código Financeiro 001. Também agradecem aos recursos do programa *Google Cloud Research Credits*, código GCP19980904.

Referências

- Cabral, B., Claro, D., and Souza, M. (2024). Exploring Open Information Extraction for Portuguese Using Large Language Models. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese*, pages 127–136.
- Carmo, D., Piau, M., Campiotti, I., Nogueira, R., and Lotufo, R. (2020). PTT5: Pretraining and validating the T5 model on Brazilian Portuguese data. *arXiv*.
- Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Burstein, J., Doran, C., and Solorio, T., editors, *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.
- Ding, N., Qin, Y., Yang, G., Wei, F., Yang, Z., Su, Y., Hu, S., Chen, Y., Chan, C.-M., Chen, W., Yi, J., Zhao, W., Wang, X., Liu, Z., Zheng, H.-T., Chen, J., Liu, Y., Tang, J., Li, J., and Sun, M. (2023). Parameter-efficient fine-tuning of large-scale pre-trained language models. *Nature Machine Intelligence*, 5(3):220–235.
- Feltrin, G., Vianna, D., and da Silva, A. (2023). Um Estudo sobre Métricas de Avaliação para Sumarização de Acórdãos. In *Anais do XXXVIII Simpósio Brasileiro de Bancos de Dados*, pages 295–305, Porto Alegre, RS, Brasil. SBC.
- Freitas, C. (2024). Dataset e corpus. In Caseli, H. M. and Nunes, M. G. V., editors, *Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português*, book chapter 13. BPLN, 2 edition.
- Fu, J., Ng, S.-K., Jiang, Z., and Liu, P. (2024). GPTScore: Evaluate as You Desire. In Duh, K., Gomez, H., and Bethard, S., editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6556–6576, Mexico City, Mexico. Association for Computational Linguistics.
- Garcia, G. L., Paiola, P. H., Morelli, L. H., Candido, G., Júnior, A. C., Jodas, D. S., Afonso, L., Guilherme, I. R., Penteadó, B. E., and Papa, J. P. (2024). Introducing Bode: A Fine-Tuned Large Language Model for Portuguese Prompt-Based Task. *arXiv preprint arXiv:2401.02909*.
- Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., Attariyan, M., and Gelly, S. (2019). Parameter-Efficient Transfer Learning for NLP. In Chaudhuri, K. and Salakhutdinov, R., editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2790–2799. PMLR.
- Hu, E. J., yelong shen, Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. (2022). LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*.
- Huang, K., Yin, H., Huang, H., and Gao, W. (2024). Towards Green AI in Fine-tuning Large Language Models via Adaptive Backpropagation. In *The Twelfth International Conference on Learning Representations*.

- Kato, M. A., Martins, A. M., and Nunes, J. (2023). *The Syntax of Portuguese*. Cambridge Syntax Guides. Cambridge University Press.
- Lacoste, A., Luccioni, A., Schmidt, V., and Dandres, T. (2019). Quantifying the Carbon Emissions of Machine Learning. *arXiv preprint arXiv:1910.09700*.
- Leal, S. E., Duran, M. S., Scarton, C. E., Hartmann, N. S., and Aluísio, S. M. (2023). NILC-Matrix: assessing the complexity of written and spoken language in Brazilian Portuguese. *Language Resources and Evaluation*, pages 1–38.
- Li, P., Yang, J., Islam, M. A., and Ren, S. (2023). Making AI less “Thirsty”: Uncovering and Addressing the Secret Water Footprint of AI models.
- Li, X. L. and Liang, P. (2021). Prefix-Tuning: Optimizing Continuous Prompts for Generation. In Zong, C., Xia, F., Li, W., and Navigli, R., editors, *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)*, pages 4582–4597, Online. Association for Computational Linguistics.
- Lin, C.-Y. (2004). ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.
- Maslej, N., Fattorini, L., Perrault, R., Parli, V., Reuel, A., Brynjolfsson, E., Etchemendy, J., Ligett, K., Lyons, T., Manyika, J., Niebles, J. C., Shoham, Y., Wald, R., and Clark, J. (2024). Artificial Intelligence Index Report 2024.
- Paes, A., Vianna, D., and Rodrigues, J. (2024). Modelos de linguagem. In Caseli, H. M. and Nunes, M. G. V., editors, *Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português*, book chapter 15. BPLN, 2 edition.
- Paiola, P. H. (2022). Sumarização abstrativa de textos em português utilizando aprendizado de máquina. Mestrado em ciências da computação, Universidade Estadual Paulista Júlio de Mesquita Filho, [s.l.]. Programa de Pós-Graduação em Ciência da Computação.
- Paiola, P. H., Garcia, G. L., Jodas, D. S., Correia, J. V. M., Sugi, L. A., and Papa, J. P. (2024). RecognaSumm: A Novel Brazilian Summarization Dataset. In Gamallo, P., Claro, D., Teixeira, A., Real, L., Garcia, M., Oliveira, H. G., and Amaro, R., editors, *Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 1*, pages 575–579, Santiago de Compostela, Galicia/Spain. Association for Computational Linguistics.
- Pontes, L., Oliveira, H., and Boldt, F. (2022). Avaliação de Modelos Neurais para Sumarização de Código-fonte. In *Anais do XLIX Seminário Integrado de Software e Hardware*, pages 140–151, Porto Alegre, RS, Brasil. SBC.
- Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1).
- Schwartz, R., Dodge, J., Smith, N. A., and Etzioni, O. (2020). Green AI. *Communications of the ACM*, 63(12):54–63.

- Souza, J. W. d. C., Cardoso, P. C. F., and Paixão, C. A. (2024). Sumarização automática. In Caseli, H. M. and Nunes, M. G. V., editors, *Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português*, book chapter 22. BPLN, 2 edition.
- Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. (2023). Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS’17, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.
- Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. (2020). Transformers: State-of-the-Art Natural Language Processing. In Liu, Q. and Schlangen, D., editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.
- Xu, L., Xie, H., Qin, S. J., Tao, X., and Wang, F. L. (2023). Parameter-Efficient Fine-Tuning Methods for Pretrained Language Models: A critical review and assessment. *CoRR*, abs/2312.12148.
- Yang, Y., Zhou, J., Wong, N., and Zhang, Z. (2024). LoRETTA: Low-Rank Economic Tensor-Train Adaptation for Ultra-Low-Parameter Fine-Tuning of Large Language Models. In Duh, K., Gomez, H., and Bethard, S., editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3161–3176, Mexico City, Mexico. Association for Computational Linguistics.
- Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M., Li, X., Lin, X. V., Mihaylov, T., Ott, M., Shleifer, S., Shuster, K., Simig, D., Koura, P. S., Sridhar, A., Wang, T., and Zettlemoyer, L. (2022). OPT: Open Pre-trained Transformer Language Models. *arXiv*.
- Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., and Artzi, Y. (2020). BERTScore: Evaluating Text Generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.
- Zhao, J., Wang, T., Abid, W., Angus, G., Garg, A., Kinnison, J., Sherstinsky, A., Molino, P., Addair, T., and Rishi, D. (2024a). LoRA Land: 310 Fine-tuned LLMs that Rival GPT-4, A Technical Report. *arXiv preprint arXiv:2405.00732*.
- Zhao, J., Zhang, Z., Chen, B., Wang, Z., Anandkumar, A., and Tian, Y. (2024b). GaLore: Memory-Efficient LLM Training by Gradient Low-Rank Projection.
- Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., et al. (2023). A survey of large language models. *arXiv preprint arXiv:2303.18223*.

TableRAG: A Novel Approach for Augmenting LLMs with Information from Retrieved Tables

Elvis A. de Souza¹, Patricia F. da Silva², Diogo Gomes²,
Vitor Batista², Evelyn Batista¹, Marco Pacheco¹

¹Applied Computational Intelligence Lab. (ICA/PUC-Rio)

²Petrobras

***Abstract.** We present TableRAG, a novel pipeline designed to integrate tabular data into traditional Retrieval-Augmented Generation (RAG) systems. Our approach is composed of three main parts: (i) generating textual representations of tables; (ii) indexing table representations in vector databases for retrieval, and (iii) employing large language models to generate SQL or Python code for data manipulation over a given table. We assessed the effectiveness of TableRAG by comparing retrieval and re-ranking accuracies over the OTT-QA benchmark and by utilizing both open and closed-source LLMs to generate code for answering questions from the WikiTableQuestions benchmark. Our best results show 86.7% HITS@5 for retrieval and 74% accuracy for Q&A, demonstrating the feasibility of integrating tabular data into RAG systems with high accuracy.*

1. Introduction

In recent years, the application of generative models in Question Answering (Q&A) systems has gained substantial traction, particularly in enterprise settings where accurate and efficient information retrieval is crucial for decision-making, customer support, and operational efficiency. Large Language Models (LLMs) have shown remarkable capabilities in generating natural language responses based on vast amounts of textual data, however, these models are not without their challenges—one of the most significant being the issue of hallucination, where the model generates plausible-sounding but incorrect or nonsensical answers [Kandpal et al. 2023, Gao et al. 2023, Lin et al. 2023, Tonmoy et al. 2024]. This problem becomes particularly pronounced when the required information is not purely textual but embedded in semi-structured data, such as tables, which can be stored in large and diversified databases, where the ability to accurately retrieve and interpret them is essential.

To mitigate hallucination and to retrieve relevant information to a particular domain, a widely adopted paradigm known as Retrieval-Augmented Generation (RAG) is employed. This technique is based on the premise that LLMs are more likely to provide accurate responses when supplied with relevant context at runtime, i.e., within the prompt that defines the task instructions, in a strategy called in-context learning [Dong et al. 2024]. Traditional Information Retrieval (IR) methods are utilized to fetch pertinent documents, which are then fed into the language model for text generation.

In this work, we propose a RAG pipeline that integrates tabular data within textual databases for information retrieval, using retrieved tables to generate reliable responses (Figure 1). Building upon traditional RAG pipelines, we implement a table retrieval module based on the same vector similarity strategy commonly used for texts, and another

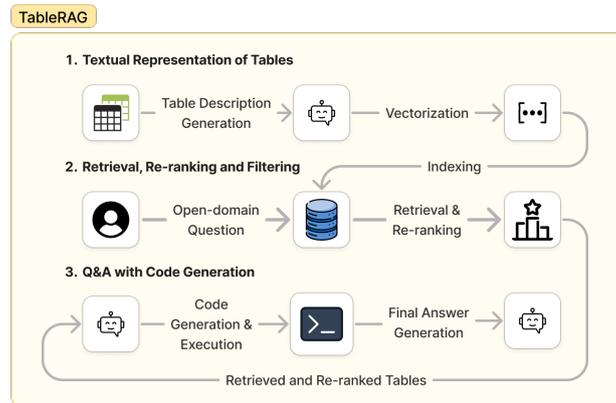


Figure 1. Overview of the TableRAG Pipeline

module for Q&A based on generating code to manipulate tables. We evaluate the results using two well-known Q&A benchmarks with tabular data, Open Table-and-Text Question Answering (OTT-QA) [Chen et al. 2020a] for the retrieval part of our pipeline, and WikiTableQuestions [Pasupat and Liang 2015] for the Q&A part.

Our results show that it is possible to incorporate tabular data into Retrieval-Augmented Generation systems in an efficient way by representing tables as texts, similarly to how traditional retrieval information systems work. Moreover, we show that code generation yields good results for obtaining answers from retrieved tables, and we highlight the considerable promise for improving the performance of open-source LLMs, which could lead to the development of more robust, adaptable, and affordable RAG systems, capable of managing multimodal data sources.

2. Related Work

Unlike Q&A over texts, which involves reading document excerpts to answer questions using extractive or generative models, Q&A over tables involves additional factors. To answer complex questions using tables, one must interpret the arrangement of rows and columns, perform filtering, joins, mathematical operations, and various other forms of table manipulation.

There are at least three tasks that have been tested in the literature for Q&A with tabular data. First, parsing semantically compositional questions in order to determine the manipulation steps that are needed to obtain a given information from a table. The WikiTableQuestions benchmark [Pasupat and Liang 2015] addresses this task. It comprises semi-structured tables, which may contain textual data, and each question may require operations such as table lookup, aggregation (counting records, summing numerical values, etc.), superlatives (finding the maximum or minimum value), arithmetic operations, among others, which need to be identified in the question in order to determine the manipulations steps that are needed to answer it correctly.

Second, manipulating relational database tables using their relationships with other tables to join information from different sources. The Spider benchmark [Yu et al. 2018] introduced this task. Its specificity involves the use of foreign keys, joining multiple tables, and constructing nested SQL queries.

Third, answering multi-hop open-domain questions that need reasoning over both textual passages and tabular information. The Open Table-and-Text Question Answering (OTT-QA) benchmark [Chen et al. 2020a] introduced this task. Its peculiarity is twofold: the retriever model must find the table that best answers the question from a large collection of tables, and the reader model must simultaneously examine data from two different modalities, texts and tables.

For the Q&A part, our approach seeks to address the challenges proposed by the WikiTableQuestions benchmark, using LLMs to generate code that correctly manipulates tables in order to find answers to semantically compositional questions. In 2015, when the WikiTableQuestions benchmark was proposed, the approach used to solve the task involved converting tables into knowledge graphs and parsing the questions into logical form, followed by selecting the most probable graph candidates to answer the question [Pasupat and Liang 2015]. The results achieved were 37.1% accuracy, considering all candidate answers, and 76.6%, considering that at least one of the candidate answers was correct.

More recent works address the WikiTableQuestions task using large language models. [Yin et al. 2020] achieved a result of 52.3% accuracy by pre-training a new model, called TaBERT, with data from 26 million tables and their respective natural language contexts [Liang et al. 2018]. The work by [Liu et al. 2024] used GPT-3.5 to transpose tables to normalize them and then used the same model to perform two types of inference: with direct prompting (DP), asking the model to reason about the table in textual form, and as a Python agent, asking the model to interact with a Python shell in up to five interactions. They achieve state-of-the-art accuracy of 73.6% on WikiTableQuestions.

3. TableRAG Pipeline Components

The instruction in Figure 2 is an example of the prompt directed to the language models for generating Python code. In this prompt, we can observe, besides the request for code generation, how the tables are represented textually and in consideration of their metadata. These procedures will be explained step by step in the following subsections.

Obtaining Textual Representation of Tables This part of our approach is inspired by the work of [Abraham et al. 2022]. Given that it is not always feasible to load entire tables into memory, the authors employ several strategies to represent the table through its schema, ensuring that the SQL query to be created is based on this indirect form of representation. The idea works particularly well in the context of Retrieval-Augmented Generation, where the initial step is to retrieve specific documents from a diverse document repository. In our case, we aim to retrieve the tables that best answer the user’s query, using the same strategy for text indexing and retrieval.

The metadata generated for a table consists of a textual description, generated by an LLM, along with additional information such as column names, unique values, and data types (numeric, textual, dates) for each table column. Other pieces of information, such as title of the table, paragraphs that explain it, acronyms meanings and so on can be inserted here to enrich the textual description of each table. When the table preview is passed to the language model, it is placed as the final element of the prompt, so to truncate it if exceeding the input token limit that the model accepts.

```

You are a backend API that is going to write a Python function named "process_df" that will return an answer to a question based
on a Pandas dataframe named "df" given as argument to the function.
Include all necessary imports inside the process_df function, including pandas and numpy.
First, you must preprocess the dataframe in order to avoid problems with type conversion.
For example, beware of missing values and preprocess unicode characters such as \xa0m if they are strings. Do not discard entire
rows.
Some columns may need to be converted into numeric types.

QUESTION: how long did it take for the new york americans to win the national cup after 1936?

This table provides information about the New York Americans soccer team. The table has six columns named Col1, Col2, Col3,
Col4, Col5, and Col6. Col1 represents the year, Col2 represents the division, Col3 represents the league, Col4 represents the
regular season performance, Col5 represents the playoffs performance, and Col6 represents the national cup performance. The
table includes data such as the team's position in the league, playoff participation, and cup achievements for different
seasons.

The column named "Col1" is about Year. Its type is object.
The column named "Col2" is about Division. Its type is float64.
The column named "Col3" is about League. Its type is object.
The column named "Col4" is about Reg. Season. Its type is object.
The column named "Col5" is about Playoffs. Its type is object.
The column named "Col6" is about National Cup. Its type is object.

The column named "Col1" has the following unique values: ['1931', 'Spring 1932', 'Fall 1932', 'Spring 1933', '1933/34',
'1934/35', '1935/36', '1936/37', '1937/38', '1938/39', '1939/40', '1940/41', '1941/42', '1942/43', '1943/44', '1944/45',
'1945/46', '1946/47', '1947/48', '1948/49', '1949/50', '1950/51', '1951/52', '1952/53', '1953/54', '1954/55', '1955/56']
The column named "Col2" has the following unique values: ['1.0', nan]
The column named "Col3" has the following unique values: ['ASL']
The column named "Col4" has the following unique values: ['6th (Fall)', '5th?', '3rd', '?', '2nd', '1st', '5th, National',
'3rd(t), National', '4th, National', '4th', '6th', '9th', '5th', '1st(t)', '8th']
The column named "Col5" has the following unique values: ['No playoff', '?', 'Champion (no playoff)', 'Did not qualify', '1st
Round', 'Finals']
The column named "Col6" has the following unique values: [nan, '1st Round', 'Final', '?', 'Champion', 'Semifinals']

Here is a preview of the table:

      Col1 Col2 Col3      Col4      Col5      Col6
0      1931  1.0 ASL      6th (Fall)  No playoff  NaN
1  Spring 1932  1.0 ASL      5th?      No playoff  1st Round
2   Fall 1932  1.0 ASL      3rd      No playoff  NaN
3  Spring 1933  1.0 ASL      ?      ?      Final
...

```

Figure 2. Prompt for Python Code Generation

Column Renaming: A common error we observed in the construction of Python code and SQL queries by LLMs is the use of non-existent columns, a particular case of hallucination. To address this issue, we implement a strategy to rename columns to temporary placeholders, such as Col1, Col2, Col3, . . . ColN. This forces language models to utilize columns based on the description of the data type they contain, rather than their names.

Table Indexing: The table is then indexed in a vector database through its textual representation. The idea is that the summary generated by the LLM should be sufficient for the table to be retrieved, leveraging the same strategy used in text indexing, by transforming table descriptions into dense contextual vectors.

Retrieval, Re-ranking and Filtering: Table retrieval is performed via similarity measures, such as cosine similarity, comparing the vector representation of the user's query with the vector representations of table descriptions. After retrieval, we employ a re-ranking step by asking an LLM to reorder the retrieved tables by relevance, and a filtering step, by asking it to filter the results, eliminating retrieved tables that despite their similarity to the user's query cannot directly answer the question.

Code Generation: Next, we prompt an LLM to generate code to obtain the answer from the table, as exemplified in Figure 2. The code generation prompt was meticulously

adjusted to avoid common pitfalls these models tend to encounter, such as attempting to perform operations on dataframe columns with missing values without first handling these cells. These adjustments were based on the outputs of GPT-4, for both SQL and Python code generation, and improved our accuracy in 16% for the same model, being the previously mentioned column renaming procedure one of the most beneficial changes we have made.

Response Generation: The generated code is then executed and the output is passed to another LLM, which is responsible for providing a natural language response to the user’s question, considering the question, the data, and the output of the executed code.

4. Methodology

To evaluate the quality of our pipeline, we utilized two traditional benchmarks, called OTT-QA [Chen et al. 2020a], for the retrieval part, and WikiTableQuestions [Pasupat and Liang 2015], for the Q&A part. The tables from both datasets were sourced from Wikipedia.

While the WikiTableQuestions dataset is ideal for our purpose of testing code generation for table manipulation, it is not ideal for evaluating the efficiency of our table retrieval module. This is because the questions are constructed in a closed-domain fashion, meaning they can only be answered if it is already known which table they refer to.¹ Therefore, when using the benchmark to evaluate our Q&A capabilities, we provide the language model with the correct table. To assess the quality of our retrieval and re-ranking modules, we will be testing them against the OTT-QA benchmark.

The OTT-QA benchmark is composed of questions based on those from another dataset, HybridQA [Chen et al. 2020b], but they were adapted to become “decontextualized,” making them open-domain and therefore suitable for testing retrieval systems. It contains 400K+ tables to retrieve from and 45K human-annotated questions. We use all the 2,214 questions in the development partition of the benchmark to test retrieval and re-ranking. We use HITS@K to measure performance, where K ranges from 1 to 5, and means whether the correct table for each question is in the top-K retrieved or re-ranked tables. Re-ranking is done after the retrieval step, and we compare results with and without re-ranking.

The WikiTableQuestions benchmark contains 2,108 semi-structured tables and 22,033 complex question-answer pairs. The questions and answers were constructed by humans through crowdsourcing, with instructions to create questions that involve various types of operations required to answer them, as shown in the distribution in Figure 3.

Due to operational constraints involving LLMs costs, all tests in the WikiTableQuestions were performed on a sample of 200 questions from the test dataset. As noted, several adjustments were made to the LLM instructions and to the table indexing method to facilitate the inference of correct answers and the inference of code that did not generate exceptions. All the adjustments were based solely on the results observed for the train

¹For example, in the question “what is the first city sorted alphabetically?”, there is no indication of which table should be found—cities of what country? in what state? in what period of time?—, and it can only be correctly answered if the table to which the question refers has already been identified.

Operation	Amount
join (table lookup)	13.5%
+ join with Next	+ 5.5%
+ aggregate (count, sum, max, ...)	+ 15.0%
+ superlative (argmax, argmin)	+ 24.5%
+ arithmetic, \cap , \cup	+ 20.5%
+ other phenomena	+ 21.0%

Figure 3. Operations required to answer a sample of 200 questions from WikiTableQuestions. Source: [Pasupat and Liang 2015].

dataset questions, thereby preventing any information leakage from test to train.

The answers in WikiTableQuestions are provided as lists of size greater than or equal to 1, with sizes greater than 1 when more than one term is expected in the predicted answer for it to be considered correct. Thus, we used accuracy to measure the performance of our Q&A module, normalizing the strings in the dataset and the strings inferred by the models, to check if all the expected terms are present in the predicted answer.²

In addition to accuracy, we also considered the number of generated codes that produced any runtime exceptions and the time taken to answer each question. In other words, we are also testing the ability of these models to generate functional and correct code in the shortest possible time.

The scenarios in which we tested our pipeline are as follows:

- For generating table descriptions, re-ranking, and the final natural language response, we consistently used the GPT 3.5 Turbo model from OpenAI, due to its good performance and low cost.
- For generating dense vector representations of each table descriptions, we use embeddings generated by a model based on XLM-RoBERTa [Conneau et al. 2019] fine-tuned for information retrieval [Wang et al. 2024], which are then indexed in the Elasticsearch platform³ and are retrieved using the cosine similarity strategy.
- For generating table manipulation code, we tested:
 - Closed-source models, namely GPT 3.5 Turbo and GPT 4 [OpenAI et al. 2024], accessed from OpenAI API;
 - Open-source models available from the GPT4All hub [Anand et al. 2023], namely LLaMa 3 8B Instruct [Dubey et al. 2024], Nous Hermes 2 Mistral DPO 7B [Jiang et al. 2023], Falcon 7B [Almazrouei et al. 2023], and GPT4All Snoozy 13B [Anand et al. 2023]. We used a single V100 32 GB GPU to infer responses with these models.
- All code generation models were tested by generating SQL queries (executed using the Python library `sqlalchemy`) and code for dataframe manipulation (executed using the Python library `pandas`).

²We use accuracy as defined in [Mallen et al. 2022], since other metrics could be harder to be applied due to the nature of generative models answers, which tend to have a lot more words than the needed terms.

³<https://www.elastic.co/>

5. Results

Figure 4 shows the performance of the retrieval and the re-ranking modules in TableRAG pipeline to find the most suitable tables to answer questions from OTT-QA. Re-ranking is applied on top of the retrieved tables, and increases the results of retrieval in 3.7 points in HITS@1, when the table ranked first is the correct one. Re-ranking is still better than just retrieval when looking at HITS@2, but after that, just retrieval becomes better than when re-ranking is applied. When looking at HITS@5, the results from retrieval are 1.4 point better than those with re-ranking. This decrease happens mainly because of lower-ranked tables that the LLM judged that needed to be filtered out of the list during re-ranking, when they were in fact the correct ones.

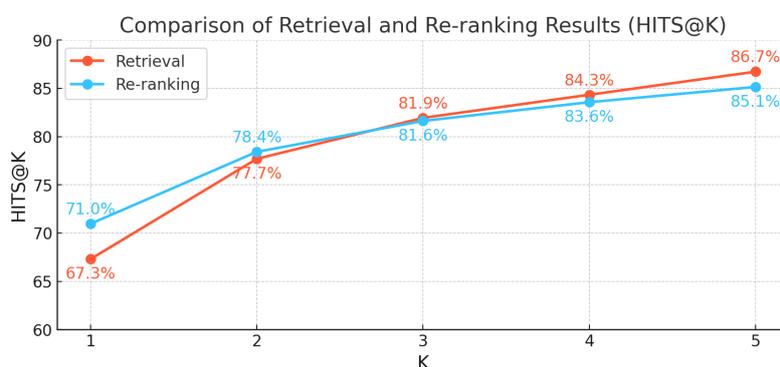


Figure 4. Comparison of Retrieval and Re-ranking Results over OTT-QA

Figure 5 shows the performance of the different models tested for generating Python code or SQL queries to answer the WikiTableQuestions. The yellow bars represent the accuracy value, while the orange ones represent the number of codes that produced exception at runtime, both in % of questions. The green line represents the average time in seconds taken to answer each question.

The best results come from OpenAI’s closed-source models, both GPT-4 and GPT-3.5. The disparity between the best closed-source model and the best open-source model can be easily explained, among other factors, by the size of the models: while Nous Hermes 2 has 7 billion parameters in its neural architecture and achieved 40% accuracy, GPT-4 is speculated to have over 1 trillion⁴, reaching 74%.

The difference in results between generating SQL and Python code for closed-source models is relatively small (74% in Python versus 72.5% in SQL for GPT-4). However, the difference in time is striking: while GPT-4 took an average of 11.5 seconds to generate Python code per question, generating SQL code was much faster, averaging 2.8 seconds. This is due to the fact that more preprocessing is required to construct functional Python code, and the larger the code, the longer the inference time. For open-source models, in general, better results are obtained when generating SQL code. Nous Hermes 2 achieved 40% accuracy building SQL queries and only 20% building Python codes, while the better open-source Python code generator, LLaMA, obtained only 26%.

⁴Sources suggest that GPT-4 could be a mixture of several models that, together, total 1.76 trillion parameters (<https://en.wikipedia.org/wiki/GPT-4>).



Figure 5. Comparison of LLMs Results in Code Generation for Q&A over WikiTableQuestions

6. Conclusions

We presented TableRAG, a pipeline for integrating tabular data into traditional Retrieval-Augmented Generation systems. The pipeline consists of obtaining textual representations of tables, indexing and retrieving them as dense contextual vectors, generating SQL or Python code for table manipulation, and generating a candidate answer in natural language. With the described pipeline, we achieved results of up to 86.7% HITS@5 in retrieval and 74% accuracy in Q&A using GPT-4.

Some limitations of this work should be considered. Due to computational cost and processing time, the results for Q&A were obtained using a sample of 200 questions from the WikiTableQuestions test set, making it difficult to compare our results with those of other works using the same benchmark. Additionally, the instructions provided to the language models for code generation were adjusted based on the outputs of the GPT-4 model, as it was the best-performing model, but they could also have been adjusted considering the most frequent errors of each of the other models individually. The open-source models did not undergo any fine-tuning process, which could significantly improve their results, and we did not test open-source models larger than 13B parameters. Finally, we did not perform any preprocessing on the tables to make them more easily interpretable, such as the table transposition procedure performed by [Liu et al. 2024], which yielded the state-of-the-art with 73.6% accuracy using GPT-3.5.

Acknowledgments

The work was carried out with assistance granted by the National Agency of Petroleum, Natural Gas and Biofuels (ANP), Brazil, associated with the investment of resources originating from the R,D&I Clauses, through the Cooperation Agreement between Petrobras and PUC-Rio.

References

- [Abraham et al. 2022] Abraham, A. N., Rahman, F., and Kaur, D. (2022). Tablequery: Querying tabular data with natural language. *arXiv preprint arXiv:2202.00454*.
- [Almazrouei et al. 2023] Almazrouei, E., Alobeidli, H., Alshamsi, A., Cappelli, A., Cojocaru, R., Debbah, M., Goffinet, E., Heslow, D., Launay, J., Malartic, Q., Noune, B., Pannier, B., and Penedo, G. (2023). Falcon-40B: an open large language model with state-of-the-art performance.
- [Anand et al. 2023] Anand, Y., Nussbaum, Z., Duderstadt, B., Schmidt, B., and Mulyar, A. (2023). Gpt4all: Training an assistant-style chatbot with large scale data distillation from gpt-3.5-turbo. <https://github.com/nomic-ai/gpt4all>.
- [Chen et al. 2020a] Chen, W., Chang, M.-W., Schlinger, E., Wang, W., and Cohen, W. W. (2020a). Open question answering over tables and text. *arXiv preprint arXiv:2010.10439*.
- [Chen et al. 2020b] Chen, W., Zha, H., Chen, Z., Xiong, W., Wang, H., and Wang, W. (2020b). Hybridqa: A dataset of multi-hop question answering over tabular and textual data. *arXiv preprint arXiv:2004.07347*.
- [Conneau et al. 2019] Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2019). Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.
- [Dong et al. 2024] Dong, Q., Li, L., Dai, D., Zheng, C., Ma, J., Li, R., Xia, H., Xu, J., Wu, Z., Chang, B., Sun, X., Li, L., and Sui, Z. (2024). A survey on in-context learning.
- [Dubey et al. 2024] Dubey, A. et al. (2024). The llama 3 herd of models.
- [Gao et al. 2023] Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., and Wang, H. (2023). Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.
- [Jiang et al. 2023] Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., de las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L. R., Lachaux, M.-A., Stock, P., Scao, T. L., Lavril, T., Wang, T., Lacroix, T., and Sayed, W. E. (2023). Mistral 7b.
- [Kandpal et al. 2023] Kandpal, N., Deng, H., Roberts, A., Wallace, E., and Raffel, C. (2023). Large language models struggle to learn long-tail knowledge. In *International Conference on Machine Learning*, pages 15696–15707. PMLR.
- [Liang et al. 2018] Liang, C., Norouzi, M., Berant, J., Le, Q. V., and Lao, N. (2018). Memory augmented policy optimization for program synthesis and semantic parsing. *Advances in Neural Information Processing Systems*, 31.
- [Lin et al. 2023] Lin, X. V., Chen, X., Chen, M., Shi, W., Lomeli, M., James, R., Rodriguez, P., Kahn, J., Szilvasy, G., Lewis, M., et al. (2023). Ra-dit: Retrieval-augmented dual instruction tuning. *arXiv preprint arXiv:2310.01352*.
- [Liu et al. 2024] Liu, T., Wang, F., and Chen, M. (2024). Rethinking tabular data understanding with large language models. In Duh, K., Gomez, H., and Bethard, S., editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long*

- Papers*), pages 450–482, Mexico City, Mexico. Association for Computational Linguistics.
- [Mallen et al. 2022] Mallen, A., Asai, A., Zhong, V., Das, R., Khashabi, D., and Hajishirzi, H. (2022). When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. *arXiv preprint arXiv:2212.10511*.
- [OpenAI et al. 2024] OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., et al. (2024). Gpt-4 technical report.
- [Pasupat and Liang 2015] Pasupat, P. and Liang, P. (2015). Compositional semantic parsing on semi-structured tables. *arXiv preprint arXiv:1508.00305*.
- [Tonmoy et al. 2024] Tonmoy, S., Zaman, S., Jain, V., Rani, A., Rawte, V., Chadha, A., and Das, A. (2024). A comprehensive survey of hallucination mitigation techniques in large language models. *arXiv preprint arXiv:2401.01313*.
- [Wang et al. 2024] Wang, L., Yang, N., Huang, X., Yang, L., Majumder, R., and Wei, F. (2024). Multilingual e5 text embeddings: A technical report. *arXiv preprint arXiv:2402.05672*.
- [Yin et al. 2020] Yin, P., Neubig, G., Yih, W.-t., and Riedel, S. (2020). TaBERT: Pretraining for joint understanding of textual and tabular data. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8413–8426, Online. Association for Computational Linguistics.
- [Yu et al. 2018] Yu, T., Zhang, R., Yang, K., Yasunaga, M., Wang, D., Li, Z., Ma, J., Li, I., Yao, Q., Roman, S., et al. (2018). Spider: A large-scale human-labeled dataset for complex and cross-domain semantic parsing and text-to-sql task. *arXiv preprint arXiv:1809.08887*.

PropBank e anotação de papéis semânticos para a língua portuguesa: O que há de novo?

Cláudia Freitas¹, Thiago Alexandre Salgueiro Pardo¹

¹Núcleo Interinstitucional de Linguística Computacional (NILC)
Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo

claudiafreitas@usp.br, taspardo@icmc.usp.br

Abstract. *This paper introduces Porttinari-base PropBank (PBP): the Porttinari-base corpus with a semantic role layer. The annotation was performed on syntactic dependencies, using linguistic rules and under human inspection. More than 40,000 arguments were annotated, and the results are discussed in light of works investigating the generalization of PropBank labels.*

Resumo. *O artigo introduz o Porttinari-base PropBank (PBP): o corpus Porttinari-base com uma camada de papéis semânticos. A anotação foi feita sobre dependências sintáticas, usando regras linguísticas e sob inspeção humana. Foram anotados mais de 40 mil argumentos, e os resultados são discutidos à luz de trabalhos que investigam a generalização das classes do PropBank.*

1. Introdução

Entre os métodos utilizados para *representar* computacionalmente informação semântica em textos está a anotação de papéis semânticos (ou SRL – *Semantic Role Labeling*). Papéis semânticos são responsáveis por indicar *quem fez o quê, para quem, onde, quando, como, por quê, para quê, com o quê, com quem*, etc., e assim estruturam de maneira *explícita e interpretável* a informação contida em enunciados linguísticos. Enquanto tarefa, a anotação de papéis semânticos tem como objetivo atribuir etiquetas a argumentos de predicadores, indicando o papel que estes argumentos exercem em uma frase.

A anotação de papéis semânticos permite criar representações semânticas *estáveis* ao longo de diferentes realizações linguísticas, e as frases 1 a 5 ilustram este ponto. Sintaticamente, “porta” exerce diferentes funções, assim como “chave”. Na atribuição de papéis semânticos, “porta” é a “coisa abrindo” em todas as 5 frases, e “chave” é o instrumento de abertura em todas as 5 frases, não importa a função sintática que exerçam. Na frase “O tempo abriu no feriado”, entretanto, teríamos uma outra estrutura argumental (e outra representação semântica), já que estaríamos diante de um outro sentido de “abrir”.

1. A chave **abriu** a porta.
2. Ela **abriu** a porta com a chave.
3. A porta foi **aberta** com a chave.
4. A porta foi **aberta** por ela com a chave.
5. A porta **abriu** com a chave.

Um PropBank (*Proposition Bank*, ou Banco de Proposições) é um corpus que contém anotação de papéis semânticos, relacionando verbos¹ e seus argumentos diretamente às estruturas sintáticas de um *treebank* e conforme o modelo sugerido por

¹Atualmente, substantivos e adjetivos também podem ser considerados.

[Palmer et al. 2005], já que, teoricamente, papéis semânticos estariam na interface sintaxe-semântica [Palmer et al. 2005, Levin 1993, Levin and Rappaport Hovav 2005].

Apesar de ser um fenômeno linguístico amplamente estudado, não há consenso relativo ao conjunto de papéis semânticos da língua. No PropBank, a diversidade teórica acerca dos papéis semânticos é contornada com a utilização de i) argumentos numerados, que vão de Arg0 a Arg5; e ii) argumentos modificadores, um conjunto mais amplo de argumentos. A motivação para o conjunto limitado de papéis numerados é facilitar a generalização para o aprendizado de máquina, ainda que alguns estudos mostrem que este objetivo é apenas parcialmente alcançado [Merlo and Van Der Plas 2009, Gung and Palmer 2021, Li et al. 2023]. A diferença entre argumentos numerados e argumentos modificadores está sobretudo na natureza da relação sintática que o argumento mantém com o verbo (*exigência*, nos argumentos numerados, vs *opcionalidade*, nos argumentos modificadores). A distinção entre os argumentos também é motivada pela sistematicidade semântica dos papéis com relação aos verbos, e por isso argumentos numerados são específicos de verbos (o Arg0 do verbo “abrir” é “quem abre”, e o Arg0 de “alagar” é “causador do alagamento”). Os ArgM, por outro lado, têm uma semântica específica e previsível (indicada pelo nome da etiqueta, como ArgM-tmp para “tempo” e ArgM-cau para “causa”), e podem se associar a qualquer verbo.

A semântica dos argumentos numerados é revelada com o alinhamento entre a anotação do corpus e o recurso lexical associado ao PropBank, os chamados *Frame Files* – em nosso caso, dispomos do Verbo-Brasil [Sanchez Duran and Aluísio 2015]. Assim, a frase (2), segundo o estilo PropBank, é anotada como indicado abaixo. Um PropBank, portanto, não é apenas um corpus anotado, mas a associação entre um corpus anotado e um léxico, que indica como os elementos devem ser anotados e o que eles significam.

- Ela[Arg0] **abriu** a porta[Arg1] com a chave[Arg2].

Neste artigo, apresentamos o *Porttinari-base PropBank* (PBP), que corresponde à adição de uma camada de papéis semânticos a todas as frases do corpus jornalístico *Porttinari-base*, que compõe o *treebank* Porttinari [Pardo et al. 2021, Duran et al. 2023]. O Porttinari-base é padrão ouro na anotação sintática conforme a teoria *Universal Dependencies* (UD) [de Marneffe et al. 2021]. Com o PBP, o Porttinari-base passa a ser duplamente padrão ouro – nas dependências sintáticas e nos papéis semânticos —, contribuindo para a disponibilização de recursos de alto nível para o processamento do português.

2. Motivação e trabalhos relacionados

O PropBank foi criado com o objetivo de treinar modelos no aprendizado supervisionado. Levando em conta a onipresença de arquiteturas neurais e grandes modelos de língua, discutimos brevemente a relevância para o PLN da anotação de papéis semânticos.

Entre as críticas ao atual paradigma de IA, estão a falta de transparência e de explicabilidade dos métodos e resultados. Neste contexto, papéis semânticos oferecem maneiras interpretáveis de representar semanticamente enunciados verbais, podendo servir de insumo, por exemplo, para a criação de grafos de conhecimento [Mohebbi et al. 2022], o que torna este tipo de representação semântica relevante para a investigação acerca da articulação entre os grandes modelos de língua (os LLMs – *Large Language Models*) e fontes de conhecimento estruturado [Dong 2023]. Na articulação entre redes neurais e

SRL, [Mohebbi et al. 2022] propõem uma abordagem de aprendizado de grafos profundos para computar similaridade semântica de documentos, usando papéis semânticos.

A anotação de SRL também pode ser usada para avaliação de modelos de língua quanto à capacidade de representar informação semântica estruturada e interpretável [Tenney et al. 2019a, Tenney et al. 2019b, Han and Choi 2020]. Nessa vertente, notamos a escassez de conjuntos de validação criados para o português, que nos faz utilizar conjuntos de dados traduzidos do inglês [Rodrigues et al. 2023]. A utilidade da anotação de papéis semânticos no PLN também pode ser indireta. Considerando o paradigma de avaliação extrínseca, [Evans and Orasan 2019] utilizam SRL para verificar se a simplificação textual é capaz de facilitar o desempenho na tarefa de SRL. Uma vez que a tarefa de SRL pode ser considerada um passo além da análise sintática, diferentes modelos de representação sintática também podem ser avaliados em função do desempenho obtido na anotação SRL, como sugerido em [Freitas 2023].

Desde 2011, o português dispõe do PropBank-BR [Duran and Aluísio 2011], que anotou papéis semânticos ao estilo PropBank sobre a parte brasileira do treebank Bosque, em sua versão de sintaxe de constituintes disponibilizada pela Linguateca. Este PropBank levou à formulação (e adaptação do inglês) de diretivas de anotação e permitiu a criação do recurso léxico que codifica os sentidos dos verbos e descreve seus frames sintáticos, o Verbo-Brasil [Sanches Duran and Aluísio 2015]. No entanto, esse recurso é ainda limitado. Tendo em vista a criação de um material maior e mais lexicalmente diversificado, foi criado o PropBank-BR v2 [Duran et al. 2014, Hartmann et al. 2016]. Diferentemente da versão anterior, este material foi construído sobre árvores sintáticas não revistas. Em ambos os casos, o processo de anotação seguiu o PropBank original, com a anotação feita de maneira linear, frase a frase. O português conta ainda com o CINTIL-PropBank [Branco et al. 2012], um corpus de frases anotadas com estrutura de constituintes e papéis semânticos, criado de maneira semiautomática, com algumas etiquetas anotadas automaticamente, e com um conjunto de papéis semânticos que é uma adaptação dos argumentos numerados de [Palmer et al. 2005]. Por fim, em uma abordagem baseada em regras, [Bick 2007] faz SRL seguindo a *Constraint Grammar*.

3. O Porttinari-base PropBank

No PBP, a anotação de papéis semânticos foi baseada em dependências, sendo cada argumento um único token. A anotação foi feita em um arquivo no formato CoNLL-U, que consiste em um texto simples com 10 campos separados por caracteres de tabulação². A anotação foi feita nas colunas 10 (nomeada de MISC) para a atribuição dos argumentos, e 9 (coluna DEPS) para a anotação dos frames. Foram anotados argumentos explícitos e implícitos, como sujeitos omitidos. A Figura 1 apresenta a codificação da frase “Júnior já presidiu a JBS, mas vendeu a sua parte”. Para facilitar a leitura, omitimos os conteúdos das colunas 3 a 6. A coluna 9 indica os frames dos verbos “presidir” e “vender” (respectivamente, presidir.01 e vender.01). A coluna 10 informa os papéis semânticos e seus predicadores. Por exemplo, o token 1, “Junior”, é Arg0 do token 3 (“presidir”) e Arg0 do token 8 (“vender”), mesmo que esta última informação não esteja explícita na frase.

A anotação do PBP utilizou 26 etiquetas. Diferentemente do Propbank original, criamos, no PBP, etiquetas que especificam alguns casos da classe mais geral ArgM-adv:

²<https://universaldependencies.org/format.html>

1	Junior	-	-	-	-	3	nsubj	-	Arg0:3 Arg0:8
2	já	-	-	-	-	3	advmod	-	ArgM-adv:3
3	presidiu	-	-	-	-	0	root	-	presidir.01
4	a	-	-	-	-	5	det	-	-
5	JBS	-	-	-	-	3	obj	-	Arg1:3
6	,	-	-	-	-	8	punct	-	-
7	mas	-	-	-	-	8	cc	-	ArgM-adv:8
8	vendeu	-	-	-	-	3	conj	vender.01	-
9	sua	-	-	-	-	10	det	-	-
10	parte	-	-	-	-	8	obj	-	Arg1:8

Figura 1. Anotação de papéis semânticos nas colunas 9 e 10 do CoNLL-U

ArgM-conseq, para indicar consequência; *ArgM-cond*, para indicar condições e *ArgM-comp*, para indicar comparações. A divergência com relação à lista de papéis do PropBank original está na classe *ArgM-src* (*source*, fonte da informação), para ocorrências como *De acordo com a polícia, trata-se de uma "prisão significativa" para as investigações*. No PropBank original, este tipo de construção não deve ser anotado, mas as anotamos pela relevância argumentativa/retórica. Em consonância com as versões do PropBank-BR, utilizamos etiquetas específicas para *verbos auxiliares* (de tempo, modo, aspecto e voz) para o pronome *-se*.

O material foi anotado com base no *Manual de anotação do PropBank-BR v2* [Duran 2014] e nos frames verbais elencados no recurso Verbo-Brasil. Ao longo do projeto, as diretivas de anotação foram enriquecidas e atualizadas, dando origem a [Duran and Freitas 2024]. Com relação a seus antecessores brasileiros, a anotação PBP difere quanto à independência da camada sintática no que se refere à identificação/segmentação de argumentos (se a segmentação da análise sintática e a segmentação de argumentos indicada no Verbo-Brasil divergirem, seguimos o Verbo-Brasil).

É interessante destacar as interferências da sintaxe UD na tarefa de SRL. No PBP, as divergências entre anotações sintática e semântica foram motivadas pela impossibilidade, em UD, de cruzar arcos sintáticos (exemplo 1), e em frases com verbos auxiliares, uma vez que a sintaxe UD é bastante econômica quanto ao que deve ser considerado verbo auxiliar. Na anotação UD do Portinari, estão anotados como auxiliares apenas auxiliares de tempo e voz. Na atribuição de papéis semânticos, porém, esta economia tem como resultado a (falsa) necessidade de atribuir papéis a elementos que, em português, não estão atuando como verbos plenos (“possam” no exemplo 2), e que portanto não deveriam receber papéis semânticos – e o resultado é uma construção sem sentido.

1. O defeito, **que** a Takata demorou a **reconhecer**, foi revelado em...
 - (a) Codificação UD: Takata demorou o defeito
 - (b) Codificação PBP: Takata reconheceu o defeito
2. O projeto **prevê** que as deduções só possam ocorrer a partir de 2021
 - (a) Codificação UD: deduções possam; prevê possam; possam ocorrer
 - (b) Codificação PBP: ocorrer deduções; prevê ocorrer

O fato de *verbos de ligação* serem considerados AUX, com relações de dependência “especiais” (os argumentos sintáticos – sujeito e predicativo – ficam dissociados do verbo “ser”, e o núcleo do sintagma é o elemento nominal predicativo) também levou à divergência de anotações, já que o verbo “ser” tem papéis semânticos para as posições de sujeito e de predicativo do sujeito.

4. Metodologia

Em termos gerais, a anotação de papéis semânticos no PBP seguiu as seguintes etapas, algumas delas concomitantes:

1. Identificação do predicador, que em nosso caso foram apenas os verbos;
2. Consulta ao Verbo-Brasil para seleção do frame adequado;
3. Anotação dos argumentos numerados conforme descritos no Verbo-Brasil;
4. Anotação dos argumentos modificadores conforme descritos nas diretivas;
5. Caso necessário, criação de frames provisórios para verbos novos ou sentidos novos de formas verbais já presentes no Verbo-Brasil;
6. Aplicação de regras de validação para detectar problemas na anotação.

Todo o processo de anotação foi feito com base em regras linguisticamente motivadas e de maneira não-linear [Wallis 2003], seguindo o exemplo de [Freitas et al. 2023]. Nisto, diferimos do processo de anotação do PropBank original, no qual cada frase era anotada inteiramente de uma vez, e do PropBank-BR versões 1 e 2, que seguiu a mesma metodologia. A anotação foi feita com a ferramenta ET [de Souza and Freitas 2021], utilizando o ambiente *Interrogatório*.

A anotação foi feita em 3 fases: (i) anotação de elementos explícitos, sempre que possível usando regras com padrões léxico-sintáticos derivados dos exemplos de Duran (2014) e das frases-exemplo no Verbo-Brasil; (ii) anotação de elementos implícitos (que envolveu sobretudo a propagação de sujeitos, feita com regras) e (iii) aplicação final de regras de validação e detecção de inconsistências. Quase todo o processo foi semi-automático, utilizando regras que associam um padrão de busca a uma regra de anotação, sempre com revisão humana. Na propagação de sujeitos omitidos de verbos na forma infinitiva, explicitamos argumentos apenas se estes fossem recuperáveis (frase 1). Em caso de dúvida ou em caso de argumentos não recuperáveis (frase 2), nada foi feito. Em [Freitas 2024] estão detalhados os procedimentos e regras utilizados.

1. O *presidente* centrista *optou* por **garantir** pela a primeira vez em anos que... (O presidente garantiu)
2. São mecânicas *que pressionam* a **entender** que isso tem custo. (não é possível determinar quem entenderá)

A anotação de elementos implícitos levou a um outro tipo de desalinhamento entre sintaxe e semântica. Na frase “A Folha **pediu** [*contato com o general Mourão*] , para que **comentasse** suas declarações, mas (...)”, o segmento “contato com o general Mourão” é Arg1 do verbo “pedir”, mas apenas “general Mourão” é sujeito/Arg0 de “comentar”.

A validação final consistiu na aplicação de 4 regras (Figura 2), que buscavam frases com condições suspeitas. Foram encontrados 101 casos suspeitos, e apenas dois deles (derivados da regra 4) eram falsos positivos. Todos os erros foram corrigidos manualmente. Diferentemente das regras utilizadas no processo de anotação, as regras de validação podem ser aplicadas para a verificação final de outros corpora com anotação de papéis semânticos. As regras de anotação, por sua vez, não foram criadas com o objetivo de serem generalizáveis para outros corpora, mas de criar um material padrão ouro de qualidade e no menor tempo possível. No entanto, a elaboração de um anotador baseado em regras, que aproveite estas regras e o corpus já anotado, é algo bastante possível.

1. Um mesmo token que contenha dois ou mais argumentos diferentes que estejam associados a um mesmo head. Esta regra garante que não há um mesmo token com papéis diferentes com relação ao mesmo verbo.
2. Dois tokens com exatamente a mesma etiqueta no que se refere a Args numerados. Esta regra garante que um verbo não terá dois Arg1 associados a ele, por exemplo.
3. Nenhum token cujo deprel é root pode ter papel semântico.
4. Não há token dependente de ArgM-mod com anotação do tipo Arg1_.*xcomp na coluna 10. Esta regra garante que todos os tokens anotados com ArgM-mod precisam ter um dependente com anotação do tipo Arg1.* na coluna 10.

Figura 2. Regras para detecção de erros

A anotação foi feita por uma única pessoa, por 7 meses, a partir das informações contidas no Manual de anotação do PropBank v2 e no Verbo-Brasil. A fim de avaliar a qualidade da anotação, foi feita uma concordância inter-anotadores *a posteriori*, tomando como base uma amostra com as 100 frases com a maior quantidade de argumentos anotados no Propbank-BR v2, sobre o qual se baseiam as diretivas de anotação e o Verbo-Brasil. Essa amostra foi reanotada sintática e semanticamente, e os resultados comparados. É importante notar também que, embora tenhamos escolhido as 100 frases com a maior quantidade de argumentos anotados, nem todos os verbos do PropBank-BR v2 têm seus argumentos anotados, apenas aqueles mais frequentes no corpus. A comparação, medida com o índice *kappa*, foi sobre 443 tokens anotados por ambas as anotações, e resultou em uma convergência de .90 (como comparação, no Propbank original, e considerando apenas a classificação de papéis incluindo Arg-M, o kappa foi de .93).

O processo de anotação dos frames dos verbos foi concomitante à anotação dos papéis semânticos. Uma vez que o foco da anotação PBP esteve na atribuição dos papéis semânticos, o processo de atribuição de sentidos aos verbos não foi exaustivo. Além de não exaustiva, a anotação privilegiou o alinhamento com verbos não monossêmicos, uma vez que a anotação de verbos monossêmicos poderia ser feita (e foi) de forma automática.

Apesar de já dispor de um recurso como o Verbo-Brasil, um corpus novo sempre traz novas formas verbais e novos sentidos para verbos já descritos. Para os sentidos (ainda) sem frames, foi feita uma busca por um verbo similar no Verbo-Brasil, e a solução foi indicada em um documento para posterior aprimoramento do Verbo-Brasil. Ao final do processo, foram documentados cerca de 350 sentidos de verbos sem frames, com exemplos do corpus e soluções provisórias em boa parte deles. A anotação de frames monossêmicos foi feita de maneira automática para os casos de verbos monossêmicos (ou seja, que só dispunham de um frame). Este procedimento levou à inclusão de mais de 500 frames no PBP.

A relação estreita entre anotação sintática e anotação de papéis semânticos, por um lado, e as interferências da anotação sintática UD, por outro, levaram à criação de diferentes versões do corpus, e com isso também criamos condições para investigar o papel de diferentes representações linguísticas no aprendizado de SRL. Cada uma das versões é gerada automaticamente a partir de uma versão base.

1. **PBP na versão UD:** Esta versão se caracteriza pela atribuição de papéis semânticos apenas aos elementos considerados verbos plenos na UD. Em consequência: (i) não foram anotados os papéis de argumentos do verbo “ser”; (ii)

foram anotados os papéis de argumentos de verbos considerados plenos pela UD, mas considerados auxiliares no Verbo-Brasil. No entanto, para diferenciá-los dos demais argumentos, receberam a etiqueta Arg1_d e Arg0_d. Se desejável, ambas as etiquetas podem ser substituídas por Arg1 e Arg0, respectivamente.

2. **PBP na versão clássica:** Esta versão prioriza o conceito de *proposição*, e se caracteriza pela atribuição de papéis conforme o modelo PropBank, independentemente do que foi considerado verbo pleno pela UD. Em consequência: (i) foram anotados os papéis de argumentos do verbo “ser”; (ii) não foram anotados os papéis de argumentos de verbos que, apesar de considerados plenos na UD, são considerados auxiliares no Verbo-Brasil e na documentação [Duran and Freitas 2024]; e (iii) foram anotados como modificadores auxiliares os verbos que, apesar de considerados plenos pela anotação UD, são considerados auxiliares no PBP (ArgM-mod, ArgM-asp), além daqueles considerados sempre auxiliares (ArgM-tml; ArgM-pas), seguindo a decisão do PropBank-BR v2 [Duran 2014].

5. Resultados e conclusões

Foram anotados 45.813 argumentos verbais e 13.395 instâncias verbais contêm anotação de frames, distribuídos em quase 1.018 frames distintos (60,8% dos verbos possui anotação de frame). As versões anteriores do PropBank-BR continham cerca de 7 mil argumentos anotados. A Figura 3 apresenta a distribuição dos argumentos por função sintática e a Figura 4 traz, com mais detalhes, os papéis semânticos mais frequentes para cada relação sintática (deprel). Todos os números se referem à versão clássica, alinhada à versão 1.0 do *trebank* Porttinari-base.

Vemos que a associação mais frequente é entre obj e Arg1, com 92,03%, seguida da associação entre nsubj e Arg0, com 63,49%. O artigo de [Palmer et al. 2005] traz o mesmo tipo de análise para o PropBank original. No entanto, uma vez que cada gramática recorta e define os elementos que lhes parecem relevantes, é necessário algum cuidado nessa comparação. O elemento sentencial/oracional S, presente no PropBank original (derivado da anotação do *Penn Treebank*), não existe em UD, e está distribuído entre alguns casos de *xcomp* e de *ccomp*, por exemplo.

De forma geral, analisando os dados do PBP, extraímos as seguintes informações: (a) Arg1 se distribui principalmente entre as relações *obj*, *nsubj*, *obl*, *xcomp* e *ccomp*; (b) Arg0 se concentra em *nsubj*; (c) Arg2 se distribui principalmente entre *obl* e *xcomp*; (d) ArgM-tmp e ArgM-mnr se distribuem entre *obl*, *advmod* e *advcl*; (e) ArgM-loc se concentra em *obl*. Com a perspectiva da sintaxe: (a) *obj* participa de Arg1 e parece ser a generalização mais fácil: “se é um obj, então é Arg1”; (b) *xcomp* participa igualmente de Arg1 e Arg2; (c) *obl* participa de Arg2, Arg1, tempo, modo e local; (d) *advmod* participa de neg, adv, tempo, dis e modo; (e) *advcl* participa de tempo, finalidade, modo e Arg2.

Os dados apontam para uma regularidade entre funções sintáticas e papéis numerados – especificamente, papéis Arg0 e Arg1. Os demais argumentos numerados, com baixa frequência, são de mais difícil generalização. Estas constatações convergem com resultados para o inglês, que verificam que a anotação ao estilo PropBank captura melhor regularidades sintáticas, sobretudo para argumentos de frequência alta, em oposição ao estilo VerNet de anotação de papéis semânticos [Merlo and Van Der Plas 2009]. Lidos de outra forma, embora o objetivo de um PropBank seja muitas vezes servir como

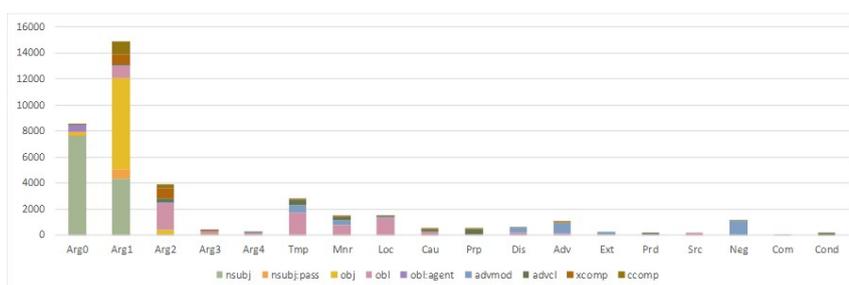


Figura 3. Distribuição de papéis semânticos por relação sintática

Deprel	Total de ocorr	Papeis semânticos mais frequentes (%)																			
NSUBJ	12038	Arg0	63,49	Arg1	35,79	Arg2	0,52	Arg3	0,17	Mnr	0,02										
OBL	6342	Arg2	31,43	Arg1	15,29	Arg0	1,31	Arg3	4,42	Tmp	26,70	Loc	20,07	Mnr	11,89	Cau	3,63	Src	3,26	Dis	2,49
OBJ	7562	Arg1	92,03	Arg0	2,26	Arg2	4,84	Arg3	0,40	Ext	0,21	Nse	0,16	Tmp	0,07						
ADVMOD	3693	Neg	29,62	Adv	20,15	Tmp	15,52	Dis	10,45	Mnr	9,83	Ext	5,58	Loc	4,82	Arg2	2,17	Arg1	1,11	Arg4	0,30
XCOMP	1790	Arg1	45,20	Arg2	45,92	Arg3	0,28	Mnr	0,45	Cau	0,89										
ADVCL	2303	Tmp	20,41	Prp	16,93	Mnr	12,77	Arg2	12,85	Cau	9,38	Cond	6,82	Prd	5,82	Adv	5,08				

Figura 4. Papéis semânticos mais frequentes para cada relação sintática

material de treino para IA, a forma de codificar os papéis talvez seja mais indicada (ou *também* seja indicada) para um anotador baseado em regras e que considere sintaxe. De fato, [Palmer et al. 2005] relatam que um anotador simples baseado em regras tem desempenho de 83%, sendo 85% o estado-da-arte em inglês, considerando cenários difíceis de avaliação, com verbos não vistos na fase de treino [Wang et al. 2022].

Apesar de raros, estudos sobre a capacidade de generalização da anotação ao estilo PropBank têm mostrado que, quando comparada à anotação ao estilo VerbNet, a anotação PropBank leva a resultados inferiores no que se refere a argumentos Arg2 a Arg5 e, em termos gerais, os bons resultados obtidos com a anotação PropBank se referem aos argumentos mais frequentes [Merlo and Van Der Plas 2009, Yi et al. 2007, Gung and Palmer 2021, Li et al. 2023]. No entanto, todos os estudos foram feitos para a língua inglesa, e apenas a disponibilização de recursos similares para o português nos permitiria verificar os resultados para a nossa língua.

Por fim, a criação de duas versões (UD e Clássica) também permite comparar a anotação PBP com outros PropBanks que tenham seguido mais fielmente a anotação UD. As diferentes versões do corpus, bem como documentação linguística detalhada e regras de anotação utilizadas, estão públicas no portal web do projeto POeTiSA³.

Agradecimentos

Este trabalho foi realizado no âmbito do Centro de Inteligência Artificial da Universidade de São Paulo (C4AI⁴), com o apoio da Fundação de Amparo à Pesquisa do Estado de São Paulo (processo FAPESP #2019/07665-4) e da IBM. Este projeto também foi apoiado pelo Ministério da Ciência, Tecnologia e Inovações, com recursos 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. Os autores agradecem também a Elvis de Souza pela preparação dos arquivos para disponibilização.

³<https://sites.google.com/icmc.usp.br/poetisa/resources-and-tools>

⁴<http://c4ai.inova.usp.br/>

Referências

- Bick, E. (2007). Automatic semantic role annotation for portuguese. In *Proceedings of TIL 2007 - 5th Workshop on Information and Human Language Technology*, pages 1713–1716, Rio de Janeiro. Sociedade Brasileira de Computação (SBC).
- Branco, A., Carvalheiro, C., Pereira, S., Silveira, S., Silva, J., Castro, S., and Graça, J. (2012). A PropBank for Portuguese: the CINTIL-PropBank. In Calzolari, N., Choukri, K., Declerck, T., Doğan, M. U., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 1516–1521, Istanbul, Turkey. European Language Resources Association (ELRA).
- de Marneffe, M.-C., Manning, C. D., Nivre, J., and Zeman, D. (2021). Universal Dependencies. *Computational linguistics*, 47(2):255–308.
- de Souza, E. and Freitas, C. (2021). ET: A workstation for querying, editing and evaluating annotated corpora. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 35–41, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Dong, X. L. (2023). Generations of knowledge graphs: The crazy ideas and the business impact. *Proc. VLDB Endow.*, 16(12):4130–4137.
- Duran, M., Lopes, L., das Graças Nunes, M., and Pardo, T. (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.
- Duran, M. S. (2014). Manual de anotação do PropBank-Br v2. Technical report, ICMC-USP.
- Duran, M. S. and Aluísio, S. M. (2011). Propbank-br: a Brazilian Portuguese corpus annotated with semantic role labels. In *Proceedings of the 8th Brazilian Symposium in Information and Human Language Technology*.
- Duran, M. S. and Freitas, C. (2024). Guia de anotação de papéis semânticos seguindo o modelo PropBank no corpus Porttinari-base. (no prelo). Technical report, ICMC-USP.
- Duran, M. S., Torres, L. S., Viviani, M. C., Hartmann, N., and Aluísio, S. M. (2014). Seleção e preparação de sentenças do corpus PLN-BR para compor o corpus de anotação de papéis semânticos Propbank-Br.v2. Technical report, Núcleo Interinstitucional de Linguística Computacional.
- Evans, R. and Orasan, C. (2019). Sentence simplification for semantic role labelling and information extraction. In Mitkov, R. and Angelova, G., editors, *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 285–294, Varna, Bulgaria. INCOMA Ltd.
- Freitas, C. (2023). Dataset e corpus. In Caseli, H. and Volpe Nunes, M. d. G., editors, *Processamento de Linguagem Natural: conceitos, técnicas e aplicações em Português*, pages 1–37. BPLN.
- Freitas, C. (2024). Anotação de papéis semânticos no corpus Porttinari-base: Procedimentos, resultados e análise. (no prelo). Technical report, ICMC-USP.

- Freitas, C., Souza, E., Castro, M. C., Cavalcanti, T., Ferreira da Silva, P., and Corrêa Cordeiro, F. (2023). Recursos linguísticos para o PLN específico de domínio: o Petrolês. *Linguamática*, 15(2):51–68.
- Gung, J. and Palmer, M. (2021). Predicate representations and polysemy in VerbNet semantic parsing. In Zarrieß, S., Bos, J., van Noord, R., and Abzianidze, L., editors, *Proceedings of the 14th International Conference on Computational Semantics (IWCS)*, pages 51–62, Groningen, The Netherlands (online). Association for Computational Linguistics.
- Han, H. and Choi, J. (2020). Establishing strong baselines for the new decade: Sequence tagging, syntactic and semantic parsing with bert. In *Proceedings of the Thirty-Third International Florida Artificial Intelligence Research Society Conference (FLAIRS 2020)*.
- Hartmann, N. S., Duran, M. S., and Aluísio, S. M. (2016). Automatic semantic role labeling on non-revised syntactic trees of journalistic texts. In Silva, J., Ribeiro, R., Quaresma, P., Adami, A., and Branco, A., editors, *Computational Processing of the Portuguese Language*, pages 202–212, Cham. Springer International Publishing.
- Levin, B. (1993). *English Verb Classes and Alternations: a preliminary investigation*. The University of Chicago Press, London.
- Levin, B. and Rappaport Hovav, M. (2005). *Argument Realization*. Cambridge University Press, Cambridge.
- Li, T., Kazeminejad, G., Brown, S., Srikumar, V., and Palmer, M. (2023). Learning semantic role labeling from compatible label sequences. In Bouamor, H., Pino, J., and Bali, K., editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15561–15572, Singapore. Association for Computational Linguistics.
- Merlo, P. and Van Der Plas, L. (2009). Abstraction and generalisation in semantic role labels: PropBank, VerbNet or both? In Su, K.-Y., Su, J., Wiebe, J., and Li, H., editors, *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 288–296, Suntec, Singapore. Association for Computational Linguistics.
- Mohebbi, M., Razavi, S. N., and Balafar, M. A. (2022). Computing semantic similarity of texts based on deep graph learning with ability to use semantic role label information. *Scientific Reports*, 12(1).
- Palmer, M., Gildea, D., and Kingsbury, P. (2005). The proposition bank: An annotated corpus of semantic roles. *Computational linguistics*, 31(1):71–106.
- Pardo, T., Duran, M., Lopes, L., Felippo, A., Roman, N., and Nunes, M. (2021). Porttinari - a large multi-genre treebank for brazilian portuguese. In *Anais do XIII Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 1–10, Porto Alegre, RS, Brasil. SBC.
- Rodrigues, J., Gomes, L., Silva, J., Branco, A., Santos, R., Cardoso, H. L., and Osório, T. (2023). Advancing neural encoding of portuguese with transformer albertina pt-*. In Moniz, N., Vale, Z., Cascalho, J., Silva, C., and Sebastião, R., editors, *Progress in Artificial Intelligence*, pages 441–453, Cham. Springer Nature Switzerland.

- Sanches Duran, M. and Aluísio, S. (2015). Automatic generation of a lexical resource to support semantic role labeling in Portuguese. In Palmer, M., Boleda, G., and Rosso, P., editors, *Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics*, pages 216–221, Denver, Colorado. Association for Computational Linguistics.
- Tenney, I., Das, D., and Pavlick, E. (2019a). BERT rediscovers the classical NLP pipeline. In Korhonen, A., Traum, D., and Màrquez, L., editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.
- Tenney, I., Xia, P., Chen, B., Wang, A., Poliak, A., McCoy, R. T., Kim, N., Durme, B. V., Bowman, S. R., Das, D., and Pavlick, E. (2019b). What do you learn from context? probing for sentence structure in contextualized word representations. In *International Conference on Learning Representations*.
- Wallis, S. (2003). Completing parsed corpora: From correction to evolution. In Abeillé, A., editor, *Treebanks: Building and Using Parsed Corpora*, pages 61–71. Springer Netherlands, Dordrecht.
- Wang, N., Li, J., Meng, Y., Sun, X., Qiu, H., Wang, Z., Wang, G., and He, J. (2022). An MRC framework for semantic role labeling. In Calzolari, N., Huang, C.-R., Kim, H., Pustejovsky, J., Wanner, L., Choi, K.-S., Ryu, P.-M., Chen, H.-H., Donatelli, L., Ji, H., Kurohashi, S., Paggio, P., Xue, N., Kim, S., Hahm, Y., He, Z., Lee, T. K., Santus, E., Bond, F., and Na, S.-H., editors, *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2188–2198, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.
- Yi, S.-t., Loper, E., and Palmer, M. (2007). Can semantic roles generalize across genres? In Sidner, C., Schultz, T., Stone, M., and Zhai, C., editors, *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 548–555, Rochester, New York. Association for Computational Linguistics.

LLMs as Tools for Evaluating Textual Coherence: A Comparative Analysis

Bryan K. S. Barbosa¹, Claudio E. C. Campelo¹

¹Systems and Computing Department
Federal University Campina Grande (UFCG) – Campina Grande, PB – Brazil

bryankhelven@ieee.org, campelo@dsc.ufcg.edu.br

Abstract. *This study evaluate recent Large Language Models (LLMs), such as GPT-4o, GPT-3.5, Claude Opus, and LLaMA 2, for their ability to analyze textual coherence. The research focuses on three areas: local coherence, where models like GPT-4o and Claude Opus excel; global coherence, where Claude Opus is most effective; and incoherence detection, where GPT-4o demonstrates strong performance. These findings reveal both the capabilities and areas for improvement in current models, shedding light on their potential applications in natural language processing, paving the way for improvements in the field.*

Resumo. *Este estudo avalia o desempenho de Grandes Modelos de Língua (LLMs) recentes, como GPT-4o, GPT-3.5, Claude Opus e LLaMA 2, na análise automática de coerência textual. A pesquisa foca em três aspectos: coerência local, onde GPT-4o e o Claude Opus se destacam; coerência global, na qual Claude Opus é o mais eficaz; e detecção de incoerências, onde GPT-4o apresenta melhor desempenho. Esses resultados revelam as capacidades e limitações dos modelos atuais, contribuindo para o entendimento de suas aplicações no âmbito do Processamento de Línguas Naturais e trazendo avanços contínuos à área.*

1. Introduction

The concept of coherence lies at the very heart of effective communication, serving as a keystone element that determines the clarity, understandability, and overall quality of textual content [Koch and Travaglia 2003]. Coherence transcends the boundaries of syntax or grammar; it embodies the logical flow of ideas, ensuring that a text is not just a collection of sentences but a unified whole that conveys meaning with precision and subtlety [Freitas 2013]. As we move deeper into the digital age, where written text interactions are increasingly prevalent [Hoey 2013], the capacity to automatically analyze textual coherence became a crucial task within the realm of Natural Language Processing (NLP).

The advent of Large Language Models (LLMs) such as GPT-3, Llama, and Gemini has revolutionized our approach to generating text that mirrors the nuance and depth of human-written content. These models, trained on extensive corpora, have demonstrated an impressive ability to produce coherent and contextually relevant text across a wide range of topics. This proficiency in text generation naturally extends to the potential for these models to excel in tasks related to textual analysis. The underlying hypothesis is simple yet profound: if an LLM can generate coherent text, it should, by extension, possess a refined ability to discern coherence – or the lack thereof – in existing texts.

In the field of computational linguistics, textual coherence is defined by the logical and orderly sequence in which ideas are presented within a text, ensuring that information and arguments are conveyed in a comprehensible and fluid manner [Seno and Rino 2005]. This involves not only the superficial connection between sentences through discourse markers or transition words but also a deeper harmony in terms of theme, purpose, and shared knowledge between the author and the reader [Charolles 1978]. For NLP systems, assessing the coherence of a text implies understanding how its constituent parts – whether at the sentence, paragraph, or document level – come together to form a unified whole that is logically consistent and aesthetically pleasing [Jurafsky and Martin 2024]. This definition highlights the complexity of the textual coherence analysis task, underscoring it as a significant challenge within the field.

Historically, coherence has been conceptualized through various theoretical frameworks. Rhetorical Structure Theory (RST) [Mann and Thompson 1987] posits that text coherence is derived from the hierarchical organization of text units, while Centering Theory [Grosz et al. 1995] emphasizes the role of discourse entities and their continuity across sentences. Over time, computational approaches to coherence have evolved from rule-based systems, relying on explicit coherence markers and structural patterns, to sophisticated machine learning algorithms that infer coherence implicitly from large datasets [Jurafsky and Martin 2024]. The development of neural network-based models, particularly those employing attention mechanisms such as BERT [Devlin et al. 2018], has marked a significant advancement, enabling a deeper understanding of contextual relationships within texts.

Given this context, the primary objective of this study is to evaluate the capabilities of various LLMs in the analysis of textual coherence. Specifically, the study assesses how the models GPT 3.5, GPT 4, GPT 4o, Claude 3 Opus, Claude 3.5 Sonnet, Claude 3 Haiku, Gemini, LLaMA 2 13b, LLaMA 2 7b, and Bard perform in three key tasks: classifying texts as (i) locally or (ii) globally coherent or incoherent, and (iii) identifying specific incoherent segments within texts. By examining these aspects, the study aims to contribute to the ongoing dialogue on improving NLP technologies and advancing our understanding of how machines process and understand the subtleties of human language.

The remainder of this article is structured as follows: Section 2 reviews key theories and models in textual coherence analysis. Section 3 presents the relevant literature, while Section 4 details the methodology, including the models evaluated and the metrics used. Section 5 discusses the results and their implications, and Section 6 concludes with a summary of findings and suggestions for future research.

2. Theoretical Background

Textual cohesion and coherence are fundamental to discourse analysis and NLP, as they explain how texts are structured and interpreted. Cohesion refers to the connections within a text created through various linguistic relations, such as pronouns, conjunctions, and lexical ties, ensuring that the text is perceived as a unified whole rather than a random collection of sentences [Halliday and Hasan 1976]. Coherence, on the other hand, is a more abstract concept, referring to the logical and meaningful organization of ideas within a text, allowing readers to follow the flow of information and understand the intended message [Van Dijk 1977].

Cohesion can be achieved through grammatical and lexical means. Grammatical cohesion includes the use of pronouns, ellipses, and conjunctions to link sentences, while lexical cohesion involves the repetition of words or the use of synonyms to maintain the continuity of ideas. However, a text can be cohesive without being coherent if the sentences do not contribute to a meaningful whole [Koch and Travaglia 2003].

Coherence can be examined at two levels: local and global. Local coherence refers to the logical connections between adjacent sentences and paragraphs, ensuring that each idea flows smoothly into the next. This is often achieved through cohesive devices, such as pronouns and conjunctions, which help maintain continuity in meaning. Global coherence, on the other hand, concerns the overall structure and unity of the text, where all parts contribute to a consistent and meaningful whole [Charolles 1978]. Both levels of coherence are essential for a text to be understood as a cohesive and logically organized entity.

Theoretical frameworks like RST and Centering Theory have been foundational in the study of coherence. For instance, RST [Mann and Thompson 1987] analyzes the hierarchical organization of text by examining the relationships between different segments, which help to structure the text in a coherent manner. Centering Theory [Grosz et al. 1995] focuses on how discourse entities are managed across sentences, ensuring that the reader can follow the progression of ideas smoothly. These models have significantly influenced NLP research, particularly in the analysis and generation of coherent texts, offering insights into the mechanisms that make a text understandable and logically connected [Jurafsky and Martin 2024].

3. Related Work

As a central area of investigation in NLP, textual coherence, particularly in the context of local coherence, which focuses on the logical and sequential flow between adjacent sentences or paragraphs, has been extensively studied through models like the entity grid. Introduced by [Lapata and Barzilay 2005] and further developed by [Barzilay and Lapata 2008], the entity grid model abstracts a text into a grid that captures the distribution and transitions of discourse entities across sentences. By analyzing these patterns, the model can effectively infer the level of local coherence within a text. This approach has been widely adopted and has inspired numerous subsequent studies. For instance, [Elsner et al. 2007] enhanced coherence assessment by integrating the entity grid with a content model, while [Lin et al. 2011] refined the method by incorporating discourse relations, further advancing the field's understanding of how sentences connect and maintain coherence.

The shuffle test, introduced by [Barzilay and Lapata 2008], has become a standard method for evaluating local coherence models. This test involves comparing the coherence of a text in its original order versus a shuffled version, challenging models to recognize the coherent sequence. Studies like those by [Lin et al. 2011] and [Dias 2016] have used this test to validate the effectiveness of their models, highlighting its importance as a benchmark in coherence evaluation.

In contrast, global coherence, which concerns the overall unity and thematic consistency of a text, has received less attention but remains a key aspect for understanding how texts function as a whole. Early work by [Thompson 1986] emphasized the role of

global coherence in enhancing readability and comprehension, arguing that a coherent text allows readers to follow the central theme or argument effortlessly. More recent contributions by [Sagi 2010] have explored the hierarchical structure of texts, demonstrating how well-organized discourse contributes to global coherence.

Recent advancements in NLP have introduced BERT and LLMs like GPT-3, which have significantly expanded the possibilities for coherence analysis. These models, trained on extensive datasets, exhibit a remarkable ability to capture both local and global coherence, leading to more refined and human-like assessments of textual structure. For example, [Braz Junior and Fileto 2021] applied BERT, specifically BERTimbau, in educational forums to measure coherence. By analyzing sentence embeddings, the model effectively assessed sentence order, accurately distinguishing between coherent and permuted texts, thus demonstrating its capability to capture nuanced textual relationships. Similarly, [Naismith et al. 2023] utilized GPT-4 for coherence assessment in educational contexts, where the model not only rated coherence but also provided explanatory rationales that closely aligned with human evaluations. This study demonstrated GPT-4's effectiveness in replicating human judgments and even surpassing traditional NLP metrics by offering rationale-supported evaluations, thereby highlighting its potential to enhance automated discourse coherence assessment and its applications in educational settings.

4. Methodology

The primary aim of this study is to analyze and compare the performance of various LLMs in evaluating textual coherence across different aspects using two distinct approaches: (i) through LLMs APIs and (ii) through LLMs chat interfaces. To achieve this, we selected a diverse set of corpora for their relevance and variety in text types, which allows for a thorough assessment of the models' capabilities across different linguistic contexts. These datasets were preprocessed and, annotated as necessary to ensure consistency across the tasks.

One of the four corpora utilized in this study is the Corpus of Contemporary American English (COCA) [Davies 2008], which offers a balanced compilation of over one billion words across genres such as spoken language, fiction, academic texts, and web pages. For our analysis, we focused on the free portions of blog and academic sections of COCA, which together comprise a diverse range of coherence levels. The blog section, with 991 texts, is characterized by its informal and subjective nature, often exhibiting lower coherence, while the academic section, consisting of 256 texts, is known for its structured and precise language, typically demonstrating higher coherence. These sections were employed in both local and global coherence tasks, with specific subsets annotated for detailed global coherence analysis and incoherence identification.

The study also incorporates the CST News Corpus [Aleixo and Pardo 2008], which consists of 50 collections of Brazilian Portuguese news articles, each centered on a specific event or topic. Originally developed to support research on multi-document summarization, the corpus includes approximately 150 news articles and 300 human-generated summaries from various newspapers, such as Folha de São Paulo, Estadão, and O Globo. This diversity in sources makes the corpus particularly well-suited for coherence studies, as it allows for evaluating both local and global coherence in a multilingual context. The CST News Corpus was especially valuable for assessing model performance

in Brazilian Portuguese, adding a multilingual dimension to our evaluations.

Another key corpus in this research is the Grammarly Corpus of Discourse Coherence (GCDC) [Lai and Tetreault 2018], which contains 4,800 texts from four real-world sources: Yahoo Answers, Clinton Emails, Enron Emails, and Yelp Reviews. Due to its context-dependent structure, the Yahoo Answers portion (1,200 texts) was excluded from our study. The Clinton Emails provide a mix of professional and personal correspondence, Enron Emails focus on formal business communication, and Yelp Reviews feature user-generated feedback on businesses. Each text is annotated for global coherence on a 3-point scale (low, medium, high), with 8,000 ratings from both expert and non-expert annotators via Amazon Mechanical Turk. These pre-existing annotations were used for comparing the models’ performance against human judgments, enhancing our evaluation of global coherence tasks.

Lastly, the DDisCo corpus [Mikkelsen et al. 2022] was developed to fill a gap in resources for studying discourse coherence in Danish. It comprises 1,002 texts from two main sources: Reddit and Danish Wikipedia. The Reddit texts, totaling 501, consist of informal user-generated content, while the 501 Danish Wikipedia texts offer more formal, structured information. Each text is annotated for global coherence on a 3-point scale (low, medium, high) by linguistics experts. This corpus introduces linguistic diversity into our research, allowing us to evaluate model performance in another non-English context. It was particularly useful for assessing how well the models generalize across different languages and discourse structures.

4.1. Local Coherence Analysis

The local coherence analysis in this study employed the shuffle test, which evaluates text coherence by comparing the original order of sentences within each text to randomly shuffled version. This test was applied to texts from four corpora: COCA, CST News, GCDC, and DDisCo. A total of 2,318 texts were selected, comprising 991 blog texts and 256 academic texts from COCA, 251 news articles from CST News, 842 texts from GCDC, and 991 texts from DDisCo. Each text was segmented into sentences, and those containing fewer than four sentences were excluded as they would not allow for the 20 required permutations. The remaining texts were shuffled 20 times, generating 46,360 incoherent versions, resulting in a dataset of 48,678 texts for analysis.

The models’ performance was evaluated using two distinct methods: (i) via LLMs APIs and (ii) through LLMs chat interfaces. In the API-based evaluation, texts were processed directly through automated API calls, streamlining the evaluation process. In contrast, the chat interface evaluation simulated real-world usage by submitting the texts through interactive prompts.

For both approaches, the models were provided with a standardized prompt for Local Coherence Analysis¹ to guide them in distinguishing between coherent and incoherent texts. Performance was measured using accuracy, precision, recall, and F1-score, comparing the models’ classifications against the original text labels.

¹<https://github.com/bryankhelven/coherence-findings>

4.2. Global Coherence Analysis

The global coherence analysis in this study aimed to evaluate the ability of various LLMs to assess the overall logical consistency and thematic organization of texts across a total of 2,142 texts. This analysis included 1,200 texts from the DDisCo Corpus and 842 texts from the GCDC Corpus, both of which already contained human annotations. Additionally, a new annotation phase was conducted for a subset of 100 texts from the COCA and CST News corpora, as these lacked pre-existing coherence labels. Three languages/linguistics experts evaluated this subset, which consisted of 10 academic texts and 60 blog texts from COCA, along with 30 news articles from CST News. For consistency, each text in this study was assigned a coherence score on a Likert scale ranging from low to high coherence (1 to 3), using the same scale previously adopted for assigning scores by the works of [Lai and Tetreault 2018] and [Mikkelsen et al. 2022]. This ensured that the evaluation of global coherence was standardized across the various corpora used in this analysis.

Following the annotation process, the study assessed the models' performance in global coherence tasks using two methods: (i) LLMs APIs for an automated process, and (ii) LLMs chat interfaces to simulate real-world, user-driven interactions. A total of 2,142 texts were used for this analysis, comprising the 100 manually annotated texts, 1,200 texts from the DDisCo corpus, and 842 texts from the GCDC corpus. Both methods utilized a standardized prompt for Global Coherence Analysis² to guide the models in assessing global coherence. The evaluation metrics were consistent with those used in the local coherence analysis, but in this case, the models' classifications were compared directly to the original human annotations (scores of 1, 2, or 3).

4.3. Incoherence Identification

The incoherence identification task evaluated the ability of various LLMs to detect segments within texts that disrupt logical flow. We used 130 texts for this task, including 100 texts previously annotated for global coherence (10 academic texts and 60 blog texts from COCA, 30 news articles from CST News) and an additional 30 texts from the GCDC corpus (10 each from Yelp, Clinton, and Enron). The same three annotators from the global coherence task identified incoherent segments, focusing on the categories of Incorrect Use of Logical Connectors, Unnecessary Repetition, Irrelevant Information, Contradictions, Sequence of Events, and Inconsistent Verb Tenses. Fleiss' Kappa, which scored 0.8326 and indicated excellent agreement, was chosen for its capacity to account for chance agreement among multiple annotators across various incoherence types. Each annotated segment was treated as a unit, ensuring robust reliability.

The annotators, familiar with each other, communicated freely to resolve difficulties, following a shared understanding of coherence from [Koch and Travaglia 2003]. The models' performance was evaluated using the same two methods as before – LLMs APIs and chat interfaces. However, in this task, each model was treated as an additional annotator. The agreement between model-generated annotations and human annotations was measured using Fleiss' Kappa to determine how closely the models aligned with human judgment. The prompt for incoherence identification³ was also standardized and used across all models in this task.

² Available on GitHub (see first footnote).

³ Ibid.

5. Results and Discussion

The results obtained during the execution of the analysis are summarized in Tables 1, 2, and 3, highlighting the performance of various models in both API and chat-based interactions.

Table 1. Performance Metrics for Local Coherence Classification

Model	API				Chat			
	Acc	Pr	Re	F1	Acc	Pr	Re	F1
Bard	0.756	0.755	0.740	0.748	0.739	0.742	0.739	0.740
Claude 3 Haiku	0.914	0.906	0.898	0.902	0.949	0.902	0.899	0.900
Claude 3 Opus	0.979	0.991	0.983	0.987	0.974	0.971	0.973	0.972
Claude 3.5 Sonnet	0.973	0.986	0.981	0.983	0.972	0.969	0.968	0.968
Gemini	0.978	0.989	0.980	0.985	0.971	0.971	0.970	0.970
GPT 3.5	0.918	0.908	0.901	0.905	0.962	0.905	0.902	0.903
GPT 4	0.970	0.982	0.980	0.981	0.969	0.966	0.965	0.965
GPT 4o	0.982	0.990	0.988	0.989	0.977	0.975	0.973	0.974
LLaMA 2 13b	0.831	0.825	0.816	0.820	0.888	0.821	0.818	0.819
LLaMA 2 7b	0.817	0.804	0.797	0.800	0.805	0.801	0.798	0.799

Table 2. Performance Metrics for Global Coherence Classification

Model	API				Chat			
	Acc	Pr	Re	F1	Acc	Pr	Re	F1
Claude 3 Haiku	0.959	0.918	0.921	0.920	0.911	0.871	0.875	0.875
Claude 3 Opus	0.982	0.986	0.987	0.986	0.933	0.936	0.939	0.937
Claude 3.5 Sonnet	0.980	0.984	0.982	0.983	0.930	0.934	0.931	0.932
Gemini	0.976	0.963	0.966	0.965	0.928	0.915	0.918	0.916
GPT 3.5	0.960	0.920	0.923	0.921	0.912	0.873	0.879	0.877
GPT 4	0.974	0.961	0.964	0.963	0.926	0.914	0.919	0.917
GPT 4o	0.978	0.965	0.968	0.967	0.930	0.918	0.920	0.919
LLaMA 2 13b	0.970	0.930	0.933	0.932	0.922	0.887	0.883	0.888
LLaMA 2 7b	0.968	0.928	0.931	0.930	0.920	0.881	0.884	0.883

Table 3. Fleiss' Kappa for Incoherence Identification

Model	API	Chat
Annotators only (baseline)	0.8326	0.8326
Claude 3 Haiku	0.7995	0.7653
Claude 3 Opus	0.8166	0.7987
Claude 3.5 Sonnet	0.8279	0.8082
Gemini	0.8119	0.7858
GPT 3.5	0.8038	0.7716
GPT 4	0.8152	0.8093
GPT 4o	0.8316	0.8234
LLaMA 2 13b	0.6787	0.6492
LLaMA 2 7b	0.5823	0.5418

Table 1 shows the performance metrics for Local Coherence Classification, with GPT 4o achieving the highest scores in both API and chat interactions. Claude 3 Opus and Claude 3.5 Sonnet also performed well, especially in the API interaction, which demonstrates their effectiveness in accurately identifying coherent texts. In contrast, LLaMA 2 13b and LLaMA 2 7b had similar lower performance on both scenarios, suggesting limitations in processing and classifying local coherence. Similarly, for Global Coherence Classification, GPT 4o and Claude 3 Opus stood out with the highest performance in both interaction modes, while Claude 3 Haiku had the lowest as shown in Table 2.

The results for the Incoherence Identification task are summarized in Table 3, where GPT 4o again demonstrated the highest agreement with human annotators, with a Fleiss' Kappa of 0.8316 in API interaction and 0.8234 in chat. Claude 3.5 Sonnet followed closely, with Kappa values of 0.8279 in API and 0.8082 in chat, while LLaMA

models, particularly LLaMA 2 7b, showed significantly lower Kappa values, indicating that these models struggle more with identifying incoherent segments.

The difference in performance between API and chat interactions is notable, with all models generally performing better in the API-based tests across all scenarios. This may indicate that API interactions allow for more precise and structured processing, leading to higher accuracy and consistency.

6. Conclusions and Future Work

This study assessed the performance of LLMs in evaluating textual coherence at both local and global levels and identifying incoherences within various corpora. Models such as GPT 4o and Claude 3 consistently outperformed others, particularly in API-based evaluations, where they achieved high accuracy and reliability. In local coherence tasks, GPT 4o demonstrated an F1 score of 0.989 in API-based tests, while in global coherence tasks, Claude 3 Opus led with an F1 score of 0.986. However, chat-based interactions revealed a performance decline, with GPT 4o's F1 score dropping to 0.974 in local coherence and Claude 3 Opus to 0.937 in global coherence. This suggests that the mode of interaction impacts model effectiveness, with API-based methods being more stable.

Despite the strong performance of top models, the Incoherence Identification task proved challenging across the board. GPT 4o showed the highest agreement with human annotators (Fleiss' Kappa of 0.8316), but all models exhibited lower performance in chat-based settings. These findings underscore the need for improvement in this area, especially as lower-tier models like LLaMA 2 struggled significantly, with Fleiss' Kappa dropping as low as 0.5418 in chat-based evaluations.

These findings have practical implications for NLP as models like GPT 4o and Claude 3 can be integrated into proofreading tools, content generators, and educational software to improve textual coherence. Their ability to assess and enhance coherence benefits machine-generated content and helps users create cohesive texts. Recognizing the impact of interaction modes on performance guides developers in choosing effective deployment strategies, favoring API integrations for consistency and accuracy.

The study acknowledges threats to validity, particularly the risk that some of the evaluation corpora may have been part of the training data for the LLMs, potentially inflating performance. This overlap introduces biases that could compromise objectivity, as models may recall patterns from training instead of genuinely evaluating coherence. The assumption of coherence in original texts and the limited size and diversity of the annotated datasets also pose risks to the generalizability of the findings.

Future work should address these limitations by expanding the range of evaluated text types and incorporating larger, more diverse annotator groups, as well as utilizing new and manually collected corpus to ensure that the models have not had prior access to it. Additionally, exploring fine-tuning techniques and evaluating newer model architectures will be essential. The development of improved evaluation metrics and the exploration of cross-linguistic and multimodal coherence analysis are also recommended to enhance the robustness and applicability of LLMs in complex language tasks.

References

- Aleixo, P. and Pardo, T. A. S. (2008). Cstnews: Um corpus de textos jornalísticos anotados segundo a teoria discursiva multidocumento cst (cross-document structure theory). Technical Report 326, Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos-SP. 12p.
- Barzilay, R. and Lapata, M. (2008). Modeling local coherence: An entity-based approach. In Knight, K., Ng, H. T., and Oflazer, K., editors, *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 141–148, Ann Arbor, Michigan. Association for Computational Linguistics.
- Braz Junior, G. and Fileto, R. (2021). Investigating coherence in posts from a doubts forum in a virtual learning environment with bert. *Conference Paper*.
- Charolles, M. (1978). *Introdução aos problemas da coerência dos textos: abordagem teórica e estudo das práticas pedagógicas*. Editora Pontes.
- Davies, M. (2008). The corpus of contemporary american english (coca). Available online at <https://www.english-corpora.org/coca/>.
- Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 1, pages 4171–4186. Association for Computational Linguistics.
- Dias, M. (2016). *Investigação de modelos de coerência local para sumários multidocumento*. PhD thesis, Universidade de São Paulo.
- Elsner, M., Austerweil, J., and Charniak, E. (2007). A unified local and global model for discourse coherence. In Sidner, C., Schultz, T., Stone, M., and Zhai, C., editors, *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 436–443, Rochester, New York. Association for Computational Linguistics.
- Freitas, A. R. P. (2013). *Análise automática de coerência usando o modelo grade de entidades para o português*. PhD thesis.
- Grosz, B. J., Joshi, A. K., and Weinstein, S. (1995). Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, 21(2):203–225.
- Halliday, M. A. K. and Hasan, R. (1976). *Cohesion in English*. Longman.
- Hoey, M. (2013). *Textual interaction: An introduction to written discourse analysis*. Routledge.
- Jurafsky, D. and Martin, J. H. (2024). *Speech and Language Processing*, chapter 23. Draft, 3 edition. Accessed: 2024-02-29.
- Koch, I. and Travaglia, L. (2003). *A coerência textual*. Editora Contexto.
- Lai, A. and Tetreault, J. (2018). Discourse coherence in the wild: A dataset evaluation and methods. In *Proceedings of SIGdial*, pages 214–223.
- Lapata, M. and Barzilay, R. (2005). Automatic evaluation of text coherence: models and representations. In *Proceedings of the 19th International Joint Conference on Artificial*

- Intelligence*, IJCAI'05, page 1085–1090, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.
- Lin, Z., Ng, H. T., and Kan, M.-Y. (2011). Automatically evaluating text coherence using discourse relations. In Lin, D., Matsumoto, Y., and Mihalcea, R., editors, *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 997–1006, Portland, Oregon, USA. Association for Computational Linguistics.
- Mann, W. C. and Thompson, S. A. (1987). Rhetorical structure theory: Description and construction of text structures. In *Natural Language Generation*, pages 85–95. Springer Netherlands.
- Mikkelsen, L. F., Kinch, O., Pedersen, A. J., and Lacroix, O. (2022). Ddisco: A discourse coherence dataset for danish. In *Proceedings of the 13th Language Resources and Evaluation Conference (LREC)*, pages 1234–1243.
- Naismith, B., Mulcaire, P., and Burstein, J. (2023). Automated evaluation of written discourse coherence using gpt-4. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 394–403, Online. Association for Computational Linguistics.
- Sagi, E. (2010). Discourse structure effects on the global coherence of texts.
- Seno, E. R. M. and Rino, L. H. M. (2005). Co-referential chaining for coherent summaries through rhetorical and linguistic modeling. In *Proceedings of the Workshop on Crossing Barriers in Text Summarization Research/RANLP*, Borovets, Bulgaria. Núcleo Interinstitucional de Lingüística Computacional – NILC/USFCAR.
- Thompson, I. (1986). Readability beyond the sentence: Global coherence and ease of comprehension. *Journal of Technical Writing and Communication*, 16(1):131–140.
- Van Dijk, T. A. (1977). Text and context: Explorations in the semantics and pragmatics of discourse.

Evaluating Federated Learning with Homomorphic Encryption for Medical Named Entity Recognition Using Compact BERT Models

Marcos F. Pontes¹, Rodrigo C. Pedrosa¹, Pedro H. Lopes¹, Eduardo J. Luz,¹

¹ Departamento de Computação – Universidade Federal de Ouro Preto (UFOP)

marcos.rezende@aluno.ufop.edu.br, {rodrigo.silva, silvap, eduluz}@ufop.edu.br

Abstract. *Medical Named Entity Recognition (NER) identifies and categorizes medical entities from unstructured texts, crucial for health monitoring tasks. Despite advancements with Large Language Models (LLMs), medical NER faces challenges due to limited and dispersed labeled data across institutions, protected under privacy regulations. Federated Learning (FL) offers a solution by enabling decentralized model training while preserving data privacy, but it is vulnerable to byzantine attacks. This research proposes a simple and secure FL protocol using Homomorphic Encryption (HE), called FedHE, that removes the need of trust between the federations and the training coordinator. Encrypted FL imposes significant constraints regarding resources consumption and performance, making the state-of-the-art language models impractical. This research aims to assess how well compact BERT representations work in federated medical NER tasks in comparison to the state-of-the-art approaches. The results showed that compact BERT representations, such as BERT_{mini} are competitive with the state-of-the-art, and are feasible to use in FedHE. However, resource consumption overheads remain a challenge, particularly when the number of clients increase.*

1. Introduction

Medical named entity recognition (NER) aims to identify medical entities (e.g., drug names, adverse reactions and symptoms) from unstructured medical texts and classify them into different categories. It can be used in many intelligent healthcare tasks such as pharmacovigilance and health monitoring [Tang et al. 2013]. With the recent advancements in the field [Peng et al. 2024] the problem of NER has seen significant improvements. However, in the specific context of medical NER, there are significant challenges in the learning process due to the sensitive nature of the data. First, the available labeled data of a single healthcare institution might not be representative enough to adjust a NER model with good predictive accuracy. Second, collaborative training with data sharing is frequently impractical considering the regulations prohibitions and the security risks associated to the data sensitiveness and trust between the parties.

To leverage massively distributed data and enhance model generalizability, federated learning (FL) was introduced in [Konečný et al. 2016] as a novel learning framework. In an FL training loop, clients collaboratively train a shared global model by exchanging model weights or gradients while keeping their data stored locally. By bringing the model to the data, FL avoids data transfer and achieves competitive performance compared to models trained with pooled data.

Recently, [Peng et al. 2024] provided an in-depth evaluation of federated learning in biomedical natural language processing, demonstrating that the BlueBERT ($BERT_{blue}$) model, particularly its larger variant ($BERT_{largeblue}$), trained using FL outperforms both its version trained on data from a single client and GPT-4 when applied in a few-shot prompt setting. Clearly, FL, combined with variations of BERT, stands out as an effective approach for NER.

Although FL’s primary focus is on maintaining rigorous privacy protections by preventing data sharing, [Zhu et al. 2019] introduced a new vulnerability in the form of inference attacks, showing that private training data can be extracted from the publicly shared gradients. To mitigate this risk, one approach is to incorporate an encryption step into the federated learning framework. Specifically, employing Homomorphic Encryption (HE) [Yi et al. 2014] within FL allows clients to encrypt their gradients, enabling the central coordinator to aggregate model updates directly on ciphertexts, thereby eliminating the need for decryption.

While HE is often considered the gold standard for data-in-use encryption, it imposes significant performance overheads in terms of both computation and communication. As a result, deploying state-of-the-art natural language models becomes impractical due to the large number of trainable parameters. Therefore, to implement a more secure FL system using HE, smaller models must be selected. In this context, this paper addresses two key questions: (i) What is the computational cost of applying HE in FL for NER applications? (ii) How much predictive accuracy might be sacrificed by choosing a more secure FL+HE approach with a smaller model? The obtained results showed that compact models, like $BERT_{mini}$, can perform competitively with state-of-the-art NER models in a FL+HE setting for different corpora. However, resource overheads — particularly communication bandwidth and memory utilization—continue to pose significant challenges.

2. Federated Learning

The prototypical FL setting consists of a central server S and a set of K distributed clients C , such that $|C| = K$, that jointly cooperate to solve a standard supervised learning task. Each client $c \in C$ has access to its own private training set $\mathcal{D}_c = \{x_{c,i}, y_{c,i}\}_{i=1}^{n_c}$. The goal of FL is to train a global predictive model whose architecture and parameters $\theta^* \in \mathfrak{R}_d$ are shared amongst all the clients and found to minimize $\min_{\theta} \sum_{c=1}^K p_c \mathcal{L}_c(\theta; \mathcal{D}_c)$, where \mathcal{L}_c is the local objective and $p_c \geq 0$ specifies the individual contribution of the client c such that $\sum_{c=1}^K p_c = 1$. Two possible configurations for p_c are $p_c = \frac{1}{K}$ or $p_c = \frac{n_c}{n}$, where $n = \sum_{c=1}^K n_c$.

The local objective function \mathcal{L}_c usually is defined as the empirical risk calculated over the training set \mathcal{D}_c sampled from the client’s local data distribution $\mathcal{L}_c(\theta; \mathcal{D}_c) = \frac{1}{n_c} \sum_{i=1}^{n_c} \mathfrak{l}(\theta; (x_{c,i}, y_{c,i}))$, where \mathfrak{l} is an instance-level loss (e.g., cross-entropy loss or squared error in the case of classification or regression tasks, respectively).

In Federated Learning, to generate a global model θ from locally trained models with parameters θ_c , an aggregation step is necessary to combine the updates from all clients. One of the most widely used methods for aggregation is called FedAvg. In each round t , clients perform local training steps on their private datasets \mathcal{D}_c to minimize their respective objective functions \mathcal{L}_c . After completing the local updates, clients send their

updated parameters $\theta_c^{(t)}$ back to the central server S . The server then aggregates these updates by computing a weighted average, typically defined as $\theta^{(t+1)} = \sum_{c=1}^K p_c \theta_c^{(t)}$, and uses this to update the global model for the next training round.

2.1. Federated Learning with Fully Homomorphic Encryption

Homomorphic Encryption (HE) allows certain computations (e.g., addition) to be performed directly on ciphertexts, without decrypting them first. The intuitive idea is that a third party can compute data without actually getting to know that data. This problem is solved with key-based encryption where the encryption process preserves algebraic operations. For the addition operator, for example, we would have $e(k, a) + e(k, b) = e(k, a + b)$ for an encryption scheme $e(\cdot, \cdot)$, encryption key k , and plaintexts a and b . A third party could thus compute the ciphertext of the value of the addition $a + b$ from the ciphertexts of a and b , and return this to the owner who could decrypt this to get the computation result on the plaintext [Al Badawi and Polyakov 2023].

In the context of FL, the viability of HE is particularly constrained when encrypting local model updates. The use of large language models, such as BERT with $110M$ parameters, becomes nearly infeasible given the bandwidth and computational overhead associated with processing encrypted gradients. This limitation underscores the need for more efficient encryption techniques, model compression strategies, and the adoption of more compact architectures—the latter being the focus of this paper’s assessment

3. Methodology

This section describes this work’s proposal for making federated medical NER secure with HE. The proposed solution, called FedHE, aims to be conceived as a generic framework on how HE can be used in FL, making them compliant with data privacy regulations and enabling scenarios such as medical NER to work without the risk of inference attacks.

The FedHE protocol uses HE encryption to protect the gradients data. Thus, even if byzantine attackers compromise the computing server, they don’t have access to the information of the gradient data from each learning client. In addition, it is impossible for byzantine attackers to use these encrypted gradient data to train shadow models.

In this work, the cryptographic scheme CKKS (Cheon-Kim-Kim-Song) [Marcolla et al. 2022] is used to encrypt clients’ gradients preserving the arithmetic operations of addition and multiplication by a scalar plaintext number. CKKS is an asymmetric cryptographic scheme that requires key pairs, so a key management service (KMS) is required. Notice that this work does not aim to detail neither the encryption scheme nor the KMS protocol, but we rely on strategies and algorithms publicly defined in the literature.

The coordinator algorithm orchestrates the federated network (See Algorithm 1). Usually, it defines how the protocol work, establish mechanisms to define the architecture, guarantee trust between the clients, and aggregate the locally generated gradients.1

The FedAvg algorithm is executed homomorphically, without decryption of clients updates. Additionally, although the KMS strategy is not specified in this paper, we assume the coordinator has only access to the public key.

Algorithm 1: FedHE Coordinator. The K clients are indexed by $c \in C$, T is the total of federated learning rounds and L is the loss function. The goal is to obtain θ^* that minimizes the clients' loss function.

State: Local model with parameters θ_i .

Function ServerTrain:

```

initialize  $\theta^{(0)}$ 
request public key from KMS
for each round  $t = 0, \dots, T$  do
  number of clients:  $m \leftarrow \max(C \cdot K, 1)$ 
  client selection:  $S_t \leftarrow$  (random set of  $m$  clients)
  for each client  $c \in S_t$  in parallel do
     $\nabla \text{enc}(L_c^{(t+1)}) \leftarrow \text{ClientTrain}(c)$ 
  end
  homomorphic FedAvg:  $\nabla \text{enc}(L^{(t+1)}) \leftarrow \sum_{c=1}^K \frac{n_c}{n} \nabla \text{enc}(L_c^{(t+1)})$ 
  for each client  $c \in C$  in parallel do
     $\text{ClientUpdate}(c, \nabla \text{enc}(L^{(t+1)}))$ 
  end
end

```

The FedHE client training algorithm is where the actual train happens (See Algorithm 2). Each client trains on their own private training dataset and share only the encrypted gradient updates with the coordinator for model aggregation.

Algorithm 2: FedHE Client. X represents the training samples while Y represents the training labels. I, ϵ, η represents the number of local epochs, the tolerance and the learning rate, respectively. The goal is to obtain θ^* that minimizes the loss function L .

State: Local model with parameters θ_i .

Function ClientTrain:

```

request public key from KMS
for each epoch  $i = 0, \dots, I$  do
  forward propagation:  $\hat{Y}_i = \text{forward}(X, \theta_i)$ 
  compute loss:  $L_i = \text{loss}(Y, \hat{Y}_i)$ 
  if  $L_i < \epsilon$  then
     $\text{break}$ 
  end
  else
    back propagation:  $\nabla L_i = \text{backprop}(X, \theta_i, L_i)$ 
    gradients encryption:  $\nabla \text{enc}(L_i) = \text{encrypt}(\nabla L_i, \text{PublicKey})$ 
    return  $\nabla \text{enc}(L_i)$ 
  end
end

```

Function ClientUpdate ($\nabla \text{enc}(L_{agg})$):

```

request private key from KMS
gradients decryption:  $\nabla L_{agg} = \text{decrypt}(\nabla \text{enc}(L_{agg}), \text{PrivateKey})$ 
update:  $\theta_{i+1} = \theta_i - \eta \nabla L_{agg}$ 

```

In conclusion, the FedHE¹ protocol is adaptable to a range of model architectures and can be seamlessly integrated into established FL platforms like Flower [Beutel et al. 2020], TensorFlow Federated², FATE³, among others.

4. Results

In this section, we present the key findings of our analysis on FedHE, focusing on two practical aspects: (1) the performance of FedHE trained with compact BERT models compared to state-of-the-art models, and (2) the performance and resource consumption overheads associated with FedHE.

Named Entity Recognition Corpora

We compared FedHE with alternative training schemes on two biomedical NLP datasets and one news dataset, focusing on NER tasks. In NER, the objective is to identify and classify named entities, such as diseases and genes, from a given sequence of tokens.

The selected corpora were chosen based on two main criteria: they are publicly available, ensuring the reproducibility of results, and they are commonly used in well-cited papers, which helps guarantee the quality of the data. A summary of the selected datasets can be found in Table 1.

Corpus	Entity/Relation Type	Corpora Type	Train	Dev	Test
CONLL-2003	General	News articles	14987	3466	3684
BC2GM	Gene	Medline abstract	26006	3251	3251
BC4CHEMD	Drug/Chem	PubMed abstract	94170	11772	11771

Table 1. List of NER corpora and their statistics

What Are the Performance and Resource Overheads of FedHE?

In FedHE, the encryption of gradients introduces a substantial increase in data size, which can significantly impact bandwidth. Table 4 provides a comparative analysis of the size overhead associated with different BERT models when using the CKKS encryption scheme. For instance, $BERT_{tiny}$, which has a plaintext gradient size of 16 MB, increases to 340 MB when encrypted. Similarly, $BERT_{mini}$'s gradient size grows from 42 MB to 864 MB under the same scheme. The most pronounced effect is seen with $BERT_{blue}$ and $BERT_{large_blue}$, where the gradient size were 20 times and more than 50 times bigger, respectively. While local training remains unaffected, these increases in ciphertext size lead to significant bandwidth overheads. As such, models like $BERT_{blue}$ become impractical for FedHE due to the prohibitive size of encrypted gradients, emphasizing the need for more bandwidth-efficient approaches or smaller models to maintain feasibility in federated settings.

Table 4 highlights the impracticality of Large Language Models in FedHE due to their exponential growth of memory and bandwidth requirements. Table 3 shows that

¹Source code: <https://github.com/marcosfpr/fedhe> and <https://github.com/marcosfpr/sealy>.

²<https://www.tensorflow.org/federated>

³<https://fate.fedai.org/>

Model	Scheme	# Params	Size
$BERT_{tiny}$	Single/Central/Federated	4M	16 MB
	FedHE	4M	340 MB
$BERT_{mini}$	Single/Central/Federated	11M	42 MB
	FedHE	11M	864 MB
$BERT_{blue}$	Single/Central/Federated	108M	415 MB
	FedHE	108M	8 GB
$BERT_{large_blue}$	Single/Central/Federated	344M	1GB
	FedHE	344M	>50 GB

Table 2. Model Parameters and Size for Different Schemes

while operations on encrypted gradients, particularly encryption, become more costly with increased parameter sizes, these do not generally pose a bottleneck in training. However, if federated training involves frequent aggregation rounds and infrequent local training epochs, these operations could become a significant bottleneck when the number of parameters is sufficiently large. Typically, clients perform extensive local training with less frequent aggregations, which aligns with both performance and operational efficiencies in FedHE.

Due to $BERT_{large_blue}$'s excessive memory demands—over 50 GB in FedHE and 1 GB in plaintext—along with significant bandwidth and processing requirements, this work will focus on comparing the effectiveness only with the base $BERT_{blue}$ model instead. Future work can be done to address the comparison with larger versions of BERT such as $BERT_{large_blue}$.

Model	Type	Mean (s)	Std. Dev. (s)	99th Percentile (s)
$BERT_{tiny}$	Encrypt	8.016	0.133	8.484
	Decrypt	2.179	0.053	2.337
$BERT_{mini}$	Encrypt	21.877	0.251	22.671
	Decrypt	5.884	0.095	6.144

Table 3. Summary Statistics for Encryption and Decryption Times of BERT Models

While Table 4 highlights significant bandwidth constraints on the server-side in FedHE, it is also essential to evaluate the resource and time costs associated with aggregation operations. Figure 4 sheds light on the performance of aggregation as the number of clients increases for the $BERT_{tiny}$ model using the $BC2GM$ corpus. The analysis indicates that the homomorphic FedAvg aggregation time does not present an efficiency issue in the training process; specifically, $BERT_{tiny}$ completes aggregation in approximately 20 seconds with 22 clients, whereas $BERT_{mini}$ requires about 50 seconds with 14 clients. However, it is important to note that as the number of clients grows, the memory required to store ciphertext gradients increases significantly. For instance, aggregation for $BERT_{mini}$ with 16 clients led to a coordinator crash due to memory insufficiency. While such issues can be mitigated using external memory strategies, these solutions introduce additional performance overhead.

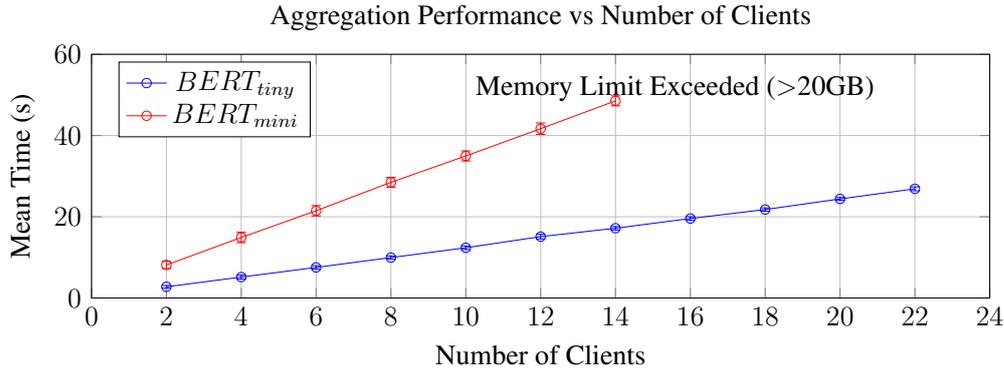


Figure 1. Aggregation performance for BERT models on BC2GM.

How Does FedHE with Compact BERT Models Compare to State-of-the-Art BERT Models For Medical NER?

Another fundamental research question for this work is to understand how far compact BERT models are from the state-of-the-art models for federated medical NER. In particular, we would also like to understand if the introduction of HE can skew the overall results. The table 4 shows an F1 comparison of the approaches tested with the state-of-the-art models.

Another critical research question addressed in this work is evaluating the performance gap between compact BERT models and state-of-the-art models for federated medical NER. Additionally, we investigate whether the integration of HE affects the overall performance outcomes. Table 4 provides a comparative analysis of F1 scores for the evaluated approaches against the state-of-the-art models.

Model	Method	CONLL-2003		BC2GM		BC4CHEMD	
		Train	Eval	Train	Eval	Train	Eval
$BERT_{tiny}$	Single	0.865 ± 0.005	0.618 ± 0.006	0.804 ± 0.002	0.593 ± 0.001	0.537 ± 0.003	0.391 ± 0.047
	Central	0.953 ± 0.001	0.728 ± 0.002	0.841 ± 0.000	0.645 ± 0.003	0.605 ± 0.011	0.460 ± 0.022
	Federated	0.624 ± 0.001	0.464 ± 0.005	0.726 ± 0.010	0.613 ± 0.016	0.598 ± 0.010	0.448 ± 0.008
	FedHE	0.816 ± 0.000	0.650 ± 0.014	0.744 ± 0.005	0.604 ± 0.010	0.621 ± 0.004	0.464 ± 0.005
$BERT_{mini}$	Single	0.961 ± 0.002	0.690 ± 0.002	0.802 ± 0.000	0.594 ± 0.002	0.802 ± 0.000	0.538 ± 0.004
	Central	0.990 ± 0.000	0.787 ± 0.005	0.995 ± 0.001	0.763 ± 0.010	0.859 ± 0.001	0.613 ± 0.002
	Federated	0.993 ± 0.000	0.758 ± 0.013	0.958 ± 0.001	0.703 ± 0.012	0.833 ± 0.000	0.584 ± 0.000
	FedHE	0.994 ± 0.000	0.781 ± 0.001	0.998 ± 0.000	0.739 ± 0.004	0.877 ± 0.000	0.590 ± 0.001
$BERT_{blue}$	Central	0.999 ± 0.000	0.791 ± 0.009	0.992 ± 0.001	0.758 ± 0.041	0.968 ± 0.001	0.683 ± 0.000

Table 4. F1 Score comparison of FedHE with various BERT models on medical NER datasets. Standard deviations are shown in parentheses. Bold indicates FedHE surpasses Federated, while underscored indicates it surpasses Centralized evaluation.

Training Setup

In all experiments we ran 50 epochs for training the models. The centralized and single-client learning, we conducted 50 local epochs on their private dataset. The single-client data were obtained splitting the dataset in two parts, and taking only one from the corpora.

The federated and FedHE approaches ran 5 aggregation rounds with 10 local epochs each on client data. Effectiveness tests, shown in Table 4, were conducted with

2 clients. Standard deviations for federated and FedHE were calculated from all clients, while single-client and centralized deviations were from 2 runs.

For training the models, we used Adam optimizer with an initial learning rate of $2e - 5$ and weight decay of 0.1. All experiments were performed on a system equipped with an NVIDIA A100 GPU and at least 32GB RAM available.

Discussion

The results in Table 4 highlight the performance of different BERT models in 4 different configurations (centralized, single client, federated and FedHE) for medical NER tasks. The centralized $BERT_{blue}$ model is used as a baseline, representing the state-of-the-art in BERT-based models for medical named entity recognition. This model sets a high standard for comparison, demonstrating its accuracy across the datasets.

Importantly, the application of HE does not skew model effectiveness. On the contrary, the encryption noise introduced by the encryption and decryption processes does not damage model accuracy. Instead, it sometimes even improves performance, as reflected in the bolded values in the table. The strongest hypothesis for this fact is that the small noise added by the ciphertext operations helped the model to generalize better. This indicates that HE can be effectively integrated without compromising, and potentially enhancing, the model’s performance.

The analysis also reveals that single-client learning models, such as $BERT_{tiny}$ and $BERT_{mini}$, often achieve higher training accuracy, but generalize with less effectively compared to federated and FedHE approaches. Federated learning and FedHE models exhibit superior generalization in all corpora evaluated.

$BERT_{tiny}$ shows lower performance compared to $BERT_{mini}$, with significant differences in all 4 methods tested for all corpora. $BERT_{mini}$, using only 11M parameters, presented satisfactory results even when compared to the more complex $BERT_{blue}$ with 108M parameters. This suggests that we can achieve results closer to state-of-the-art using compact BERT representations without making FL+HE impractical. This work also suggests that evaluating other slightly more complex BERT variants, such as $BERT_{small}$, could provide additional insights and potential improvements in model performance.

5. Conclusion

Overall, FedHE shows a generic framework for integrating HE in a FL protocol as a strong alternative for federated medical NER tasks. FedHE offers robust performance and practical advantages, making it a compelling choice for scenarios where data privacy and model effectiveness are critical. The results underscore the viability of FedHE in maintaining high performance while incorporating encryption techniques. For future work, we highlight (1) study scenarios where the number of clients is higher and the data is non-IID; (2) assess the feasibility of other compact BERT variants such as $BERT_{small}$ and (3) test the models against other LLM-based baselines such as $BERT_{large.blue}$.

6. Acknowledgements

The authors would also like to thank the Universidade Federal de Ouro Preto (PROPPI/UFOP) for supporting the development of this study.

References

- Al Badawi, A. and Polyakov, Y. (2023). Demystifying bootstrapping in fully homomorphic encryption. *Cryptology ePrint Archive*.
- Beutel, D. J., Topal, T., Mathur, A., Qiu, X., Fernandez-Marques, J., Gao, Y., Sani, L., Li, K. H., Parcollet, T., de Gusmão, P. P. B., et al. (2020). Flower: A friendly federated learning research framework. *arXiv preprint arXiv:2007.14390*.
- Konečný, J., McMahan, H. B., Yu, F. X., Richtárik, P., Suresh, A. T., and Bacon, D. (2016). Federated learning: Strategies for improving communication efficiency. *arXiv preprint arXiv:1610.05492*.
- Marcolla, C., Sucasas, V., Manzano, M., Bassoli, R., Fitzek, F. H., and Aaraj, N. (2022). Survey on fully homomorphic encryption, theory, and applications. *Proceedings of the IEEE*, 110(10):1572–1609.
- Peng, L., Luo, G., Zhou, S., Chen, J., Xu, Z., Sun, J., and Zhang, R. (2024). An in-depth evaluation of federated learning on biomedical natural language processing for information extraction. *npj Digital Medicine*, 7(1):127.
- Tang, B., Cao, H., Wu, Y., Jiang, M., and Xu, H. (2013). Recognizing clinical entities in hospital discharge summaries using structural support vector machines with word representation features. In *BMC medical informatics and decision making*, volume 13, pages 1–10. Springer.
- Yi, X., Paulet, R., Bertino, E., Yi, X., Paulet, R., and Bertino, E. (2014). *Homomorphic encryption*. Springer.
- Zhu, L., Liu, Z., and Han, S. (2019). Deep leakage from gradients. *Advances in neural information processing systems*, 32.

A Dependency Treebank of Tweets in Brazilian Portuguese: Syntactic Annotation Issues and Approach

Ariani Di Felippo^{1,2}, Maria das Graças V. Nunes¹, Bryan K. da Silva Barbosa^{1,3}

¹Núcleo Interinstitucional de Linguística Computacional (NILC)

²Departamento de Letras, Universidade Federal de São Carlos (UFSCar)
Caixa Postal 676 -- 13565-905 -- São Carlos -- SP -- Brazil

³Programa de Pós-Grad. em Linguística, Universidade Federal de São Carlos (UFSCar)
Caixa Postal 676 -- 13565-905 -- São Carlos -- SP -- Brazil

ariani@ufscar.br, gracian@icmc.usp.br, bryan42@estudante.ufscar.br

Abstract. We broaden Brazilian Portuguese (BP) dependency parsing to handle “user-generated content” by developing and annotating the first BP treebank of tweets (actual X posts) within the Universal Dependencies framework. DANTEStocks has a size of 4,048 tweets from the stock market domain already annotated with PoS tags and morphological features from UD. In this paper, we describe our standards for dealing with Twitter- and domain-specific properties of the corpus in the dependency annotation process. The enriched version of DANTEStocks with dependency relations from UD and the annotation guidelines are already publicly available.

Resumo. Amplia-se a análise de dependência do português brasileiro (pt-br) para lidar com “conteúdo-gerado por usuários” ao desenvolver e anotar o primeiro treebank de tweets (atuais posts do X) em pt-br segundo o modelo Universal Dependencies. O DANTEStocks possui 4,048 tweets do mercado financeiro e anotação-UD de tags PoS e traços morfológicos. Neste artigo, descreve-se a estratégia de anotação sintática adotada para lidar com as idiosincrasias do Twitter e do domínio desse corpus. A versão do DANTEStocks enriquecida com as relações de dependência-UD e as diretrizes de anotação já estão publicamente disponíveis.

1. Introduction

The *Universal Dependencies* (UD) [Nivre *et al.* 2020] project specifies a complete morphological and syntactic representation with the goal of facilitating multilingual tagger and parser development [Nivre 2016]. The morphology of a word consists of 3 levels of information: PoS tag, lemma, and features. Syntactic annotation consists of typed dependency relations (*deprels*) between words. Currently, the model has 17 PoS tags and 37 *deprels*, plus a non-fixed set of morphological features. Figure 1 shows an example of an annotated tweet in DANTEStocks. In a dependency tree, one word is the head of the utterance (`root`) and all other words are dependent on another word. The labeled arcs represent the *deprels*, pointing from heads to their dependents. The PoS tag and the lemma of each word are displayed below the text. The morphological features are not included in this figure. However, the token “acordo” (“agreement”), for example, has the following *features* and values according to UD: `Number=Sing, Gender=Masc.`

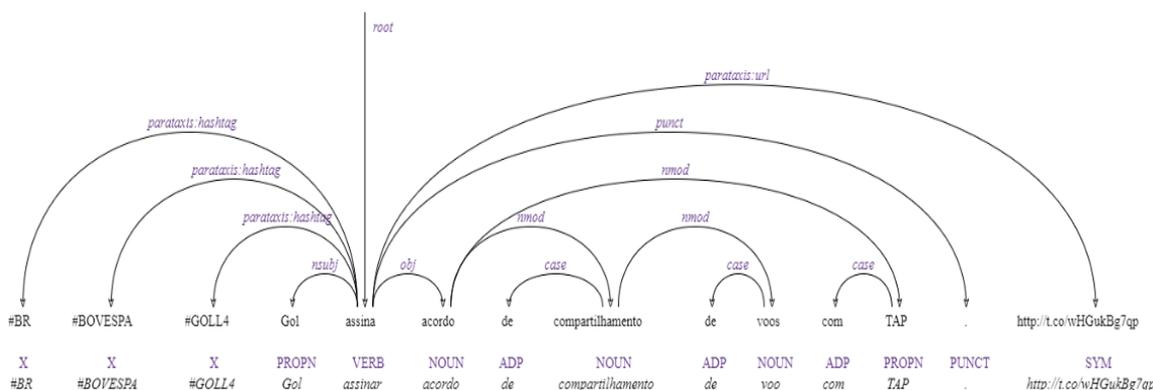


Figure 1. UD annotation of “#BR #BOVESPA #GOLL4 Gol assina acordo de compartilhamento de voos com TAP - <http://t.co/wHGukBg7qp>”¹.

Motivated by UD, treebanks for new domains, genres and language varieties have been recently built. Among the treebanks featuring *user-generated content* (UGC) created from 2014 onwards, a significant number is either partially or entirely made up of Twitter data, whose language diverges from standard written texts in several ways, posing significant challenges for building UD-based treebanks. These challenges include non-standard spelling, capitalization, punctuation, syntax, platform conventions, and creative language use, which often introduce many unknown words. Promoting cross-linguistic consistency, UD guidelines for UGC annotation have been provided [e.g. Sanguinetti *et al.* 2022], however, when it comes to a technical domain, specific strategies are required. Due to the variety and complexity of the language, adequate treatment of the phenomena by means of an already existing model, such as UD, is a non-trivial task.

We report the syntactic annotation of DANTEStocks within UD framework. First, we briefly describe the segmentation, tokenization, and the previous PoS annotation of the corpus (§2). Then, we present the annotation guidelines for the UD-*deprels* (§3). In (§4), we detail the semi-automatic approach for annotating the dependency relations, including data organization, creation of a reference subcorpus, and training a state-of-art parsing model on tweets. In (§5), we report a small-scale evaluation of the syntactic annotation. Finally, we put our work into context and outline future work (§6).

2. The DANTEStocks Corpus

DANTEStocks is a corpus comprising 4,048 tweets (with 140-character limit) from the stock market domain. It was automatically collected by fetching posts containing a *ticker*² of one of the 73 stocks that compose the Ibovespa³. Considering the entire tweet as a basic unit for syntactic analysis, the DANTEStocks’ tweets are not segmented into smaller units (sentences, clauses or phrases). This decision saved the effort to conduct a manual segmentation or do revision of an automatic process. Additionally, the corpus was not normalized to preserve its diversity, as the goal was to develop multigenre applications. Although focusing on syntax, we outline the previous segmentation and morphological UD-annotation because they contextualize some annotation decisions.

¹ “#BR #BOVESPA #GOLL4 Gol signs flight sharing agreement with TAP - <http://t.co/wHGukBg7qp>”

² It is a five or six-character alphanumeric string that represents a specific type of stock from a company, such as “PETR4” for Petrobras’ preferred stock.

³ It is the benchmark indicator of B3 (“*Brasil, Bolsa, Balcão*”), which is the main financial exchange in Brazil.

Following the lexicalist view of syntax of UD, the syntactic words⁴ (tokens) were automatically segmented by a version of the NLTK TweetTokenizer⁵, augmented with specific rules for UGC [Silva *et al.* 2021]. The tool preserves most white-space-delimited tokens, including phonetization (e.g. “d+” > “demais”), hashtag, cashtag⁶, at-mention, emoticon, and URL, and splits off single orthographic tokens that correspond to multiple (syntactic) words, such as clitics, contractions (canonical and non-canonical), punctuation marks (except for abbreviations), and valuation rates and monetary values with unconventional orthography. After the manual revision of the tool output, the corpus ends up with a total number of 81,037 tokens.

The morphological annotation was also conducted semi-automatically⁷ [Silva *et al.* 2021]. The PoS tags generated by the UDPipe 2 parser [Straka 2018], trained incrementally over UD-Portuguese Bosque [Rademaker *et al.* 2017] and tweets, were manually analyzed by three annotators, and the cases of disagreement among them were adjudicated by a senior linguist based on guidelines tailored for standard texts in BP [Duran 2021] and tweets [Di-Felippo *et al.* 2022]. All 17 UD-tags can be found in DANTEStocks. PUNCT, NOUN, and PROP are the most frequent, with around 16%, 15% and 14% of all the tags, respectively. Lemmas and grammatical features were semi-automatically obtained by using the PortiLexicon-UD lexicon [Lopes *et al.* 2022]. Major manual adjustments were required for lemmatization due to the high rate of out-of-vocabulary words. Regarding grammatical features, the scenario was quite different. The features extraction was guided by the already validated PoS tags and lemmas, which decreased the manual revision effort. Most of the corrections was related to errors arising from ambiguity about VERB class features (*VerbForm*, *Mood*, *Tense*, *Gender*, *Number* and *Person*). The manual revision also focused on checking *Typo*, *Abbr*, and *Foreign*, which are features that can be associated to words belonging to all PoS classes.

While many syntactic structures of tweets could be quite straightforwardly annotated using the general guidelines adapted for Portuguese [Duran 2022], many of them needed specific choices. In the next section, we discuss the main challenging issues for annotation decisions related to dependency relations (*deprels*).

3. Syntactic Annotation Issues

3.1. Medium- and domain-dependent (lexical) phenomena

Mostly following the recommendations of Sanguinetti *et al.*, tokens classified as orthographic variation from standard norm by [Scandarolli *et al.* 2023] were annotated with their actual syntactic roles, since they are always syntactically integrated. These variations include user-generated content phenomena such as substitution, omission, insertion, and transposition of characters (e.g., letters, spaces, hyphens, and diacritics). A good example is the token “*nao*” (instead of “*não*”) (“no”) in (1) “*VALE5 nao passa de 29,9*”⁸, which has a case of diacritic omission. In the example, “*nao*” was related to the root “*passa*” by *advmod*, since it is an adverb that modifies a predicate.

⁴ It is the basic annotation unit that plays a syntactic function in an utterance.

⁵ <https://www.nltk.org/api/nltk.tokenize.html>

⁶ It was specifically designed to track financial instruments (e.g., \$PETR4).

⁷ The version of the corpus containing PoS and features annotation is publicly available at: <https://sites.google.com/icmc.usp.br/poetisa/resources-and-tools>.

⁸ “VALE5 does not exceed 29,9”

The same strategy was adopted for treating most of the phenomena classified as “innovative norm”⁹ by [Scandarolli *et al.* 2023] (*i.e.*, abbreviation, neologism, mark of expressiveness and homophone writing), since they are also always syntactically integrated. Pictogram (emoticon/emoji), which is a mark of expressiveness, is the only one that occurs non-syntactically integrated (standalone), being attached to the **root** by *discourse*. The other two types of innovative norm’s phenomena required annotation guidelines when standalone and syntactically integrated (Table 1). For the medium-dependent devices, the treatment given to the at-mentions when preceded by the RT mark is only that differs from the recommendation of Sanguinetti *et al.* Instead of considering the at-mention as standalone and attaching it to the main predicate with *vocative*, we treat it as a syntactically integrated token attached to the RT mark by *nmod*. This is due to our interpretation of an elliptical preposition “*de*” (“of”) (“RT **de** @user”), indicating an attributive relationship between the RT/SYM and the @user/PROPN. Also differently, all the cases of *parataxis* involving a UGC phenomenon in DANTEStocks are annotated with a corresponding subrelation, not only for URL and hashtags.

Table 1. UD-dependency guidelines for Twitter- and domain-specific issues.

UGC issue	Subtype	Syntactic integration	Standard syntactic role	Other
Medium-dependent token	<i>Hashtag</i>	No		<i>parataxis:hashtag</i>
		Yes	✓	
	<i>At-mention</i>	No		<i>parataxis:mention</i>
		Yes	✓	<i>nmod</i> (of the RT)
	<i>URL</i>	No		<i>parataxis:url</i>
		Yes	✓	
	<i>RT</i>	No		<i>parataxis:rt</i>
		Yes	✓	
<i>Truncation</i>	Yes	✓	(<i>:wtrunc</i> or <i>:strunc</i>)	
<i>Code-switching</i> (intra)	Yes	✓	(if known)	<i>flat:foreign</i> (if unknown)
Domain-specific token	<i>Ticker</i>	Yes	✓	
	<i>Cashtag</i>	No		<i>parataxis:cashtag</i>
		Yes	✓	

3.2. Unconventional syntax

Besides all the linguistic issues previously mentioned, the complexity of the UD-annotation also rises from the highly contextual nature of Twitter, and the high level of fragmentation that seems to be typical in UGC from stock market domain. This provides a rich context for ambiguities and ellipses, resulting in unconventional syntactic structures whose most appropriate UD analysis depends on the interpretation of the tweet content. One example is *nsbuj:pass* without the *aux:pass*. To recommend attaching “#cyre3” to the *root* “*postado*” by *nsbuj:pass* in the tweet of Figure 2, we assumed that the auxiliary verb is elided. In Figure 2, we also assumed an elliptical preposition (“*a*”) preceding “+1,78” to connect “1,78” by *obl*. Since the syntactic function of “(+)*1,78*” is ambiguous (*i.e.*, *obl* of “*postado*” or *nmod* of “*abertura*”), the choice of “1,78” as dependent on the *root* by *obl* illustrates annotation decisions based on the interpretation of domain experts.

⁹ They are lexical alternatives to existing standard words and frequent linguistic devices that are found in the Twitter and/or stock market domain language [Scandarolli *et al.* 2023].

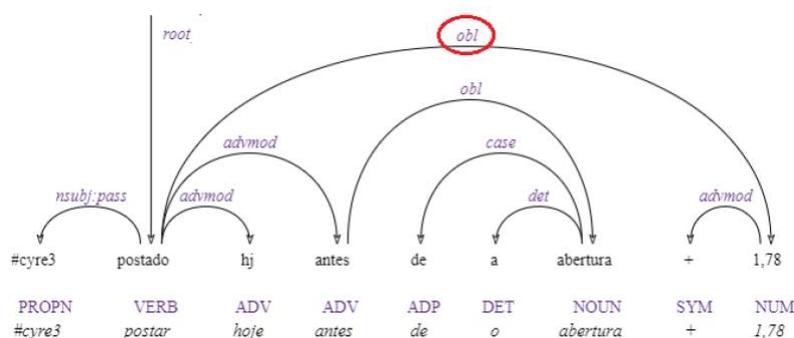


Figure 2. Syntactic ellipsis in the fragment ““#cyre3 postado hj antes da abertura +1,78””¹⁰.

3.3. Structural patterns

Besides the UGC (lexical) phenomena and unconventional syntax issues, we also identified 22 recurring structural patterns among the tweets in DANTEStocks. Such patterns correspond to almost 1,000 instances of the corpus, *i.e.* unique tweets. For each pattern, we created a template for guiding the annotation of the pattern instances in the corpus. The 22 templates also compose the dependency annotation guidelines for the DANTEStocks corpus, as well as the recommendations for the treatment of the lexical phenomena and unconventional structures [Di-Felippo *et al.* 2024].

More precisely, a template contains 3 fields: (i) pattern, *i.e.* a mnemonic description, (ii) elements, *i.e.* list of pattern elements and the corresponding annotation guideline within UD, and (iii) example, *i.e.*, at least one attested instance of the pattern from the corpus with its UD-dependency annotation. It is important to mention that, since the patterns usually refer to fragmented and/or full of syntactic ellipsis tweets, the template specification is based on a possible interpretation of the tweets, which was done with the support of stock market’s experts.

For illustration, the Template 11 is shown in Table 2. It corresponds to 20 unique instances in the corpus. Since the pattern of the template represents very fragmented tweets, the domain experts helped us to interpret corpus utterance such as that in Table 2 as being composed by three blocks of information, resulting in the following pattern description: <hashtag-ticker><theme><url>.

The <theme> provides information about a specific stock, codified by the <hashtag-ticker>, and it was considered the main information of the utterance. Since the <theme> is always being introduced by the coordinate expression “support and resistance”¹¹, the first element of the expression (*i.e.* “suportes”) is the `root`, as indicated in the field “elements”. In the “element” field, it is also indicated that the <hashtag-ticker> is dependent on the `root` with the `nmod` tag, due to interpretation of “#VALE” as a nominal that functionally corresponds to a modifier of another noun (“*suportes*”). Since the `nmod` relation is usually introduced by a preposition (ADP tag) in Portuguese, we assume, to propose the template, that there is an elliptical preposition “*de(+a)*” (*i.e.* “*suportes e resistências da VALE4*”) (“support and resistance” of #VALE5). Finally, the <url> is dependent on the `root` with `parataxis:url` because it is a run-on segment.

¹⁰ “#cyre3 posted today before opening +1,78”.

¹¹ Terms that indicate price levels where a specific stock tends to reject the current trend and reverse, *i.e.*, they indicate potential turning points in a stock’s price.

Table 2 Template for UD-dependency annotation of tweets with structural pattern.

Pattern	<hashtag-ticker> <theme> <url>, where:
Elements	a. <hashtag-ticker> is dependent on the root with the nmod label b. <theme> contains the expression “ <i>suportes e resistências</i> ”; “ <i>suportes</i> ” is the root c. <url> is dependent of the root with the parataxis:url tag
Example	#VALE5 suportes e resistências http://t.co/c8OrWXrECN

4. Syntactic Annotation Approach

The dependency-based annotation of DANTEStocks was held in two semi-automatic stages [Barbosa 2024]. The first one aimed at creating a reference subcorpus and the second stage of the annotation focused on fine-tuning a pre-trained parser for tweets by using the reference subcorpus as part of its initial training set. To start the syntactic annotation, all 4,048 tweets were grouped into three major sets, capturing tweets with: (i) relatively standard language, (ii) recurring structural patterns, and (iii) other (tweets that do not belong to any of the first two sets). Tweets were classified through *k-means* clustering [Macqueen 1967] with *tf-idf* (“term frequency–inverse document frequency”) [Luhn 1957].

4.1. Creation of a Reference Subcorpus

The organization of tweets into sets as mentioned above allowed us to select a few instances from each set, covering all the lexical and structural diversity of DANTEStocks to compose a reference subcorpus of 1,000 tweets. Furthermore, as an attempt to achieve annotation consistency, particularly given the non-canonical language of the corpus, the semi-automatic annotation of the subcorpus was also based on such classification. This means that the data from each major set was manually reviewed separately.

To create a gold-standard subcorpus we also used the UDPipe 2 parser trained over UD-Portuguese Bosque to annotate the 1,000 tweets. The UD-annotated subcorpus was later manually revised by a single expert. Taking advantage of the previous experience of the expert in UD-annotation of journalistic texts and the training of UDPipe 2 over Bosque, the manual revision started with tweets that present relatively standard language. The next tweets were those with recurring structural patterns, and finally the tweets with a variety of lexical and structural characteristics. During the revision process, the challenging issues described in Section 3 were discussed, and the annotation decisions gave rise to the guidelines for the treatment of tweets from the stock market domain within the UD framework [Di Felippo *et al.* 2024]. The guidelines were used to support the manual revision of the rest of the corpus, which was done when training a state-of-art parser on the tweets from DANTEStocks. After the revision of the subcorpus, we ended up having a gold-standard subset of 1,000 syntactically annotated tweets.

4.2. Parsing model training

The rest of the corpus was annotated by customizing Stanza [Qi *et al.* 2020] for DANTEStocks. Stanza is a well-known pre-trained model for Portuguese, having the advantage of being a user-friendly pipeline for text analysis. The procedure began with the Stanza base architecture, fine-tuned on Porttinari-base [Duran *et al.* 2023], which is a journalistic corpus composed of 8,418 sentences (168,080 tokens) manually annotated with UD, and the reference subcorpus. For the first run of Stanza, comprising Porttinari-base and the reference subcorpus as initial training dataset, was applied the same distribution of data found in Porttinari-base¹², resulting in a dataset of 9,893 samples, being 70% for training, 10% for validation, and 20% for testing. The resulting parser was used to annotate a new package of data (out the remaining 3,048 tweets), which was manually revised and incorporated to the previous data set, being then used to start a new training run of Stanza. This cycle continued incrementally until the last package of tweets was annotated/revised. Besides the first training iteration, we carried out five training runs, adding packages of 203, 300, 400, 400, and 1233 tweets per iteration, respectively (totaling 2,536 tweets). The resulting model of the 6th (final) run was used to annotate the remaining 512 tweets. The tweet packages were added in the same order as the manual revision of the reference subcorpus: standard language tweets, structural pattern tweets, and tweets with varied lexical/structural properties.

For each of the five runs, we kept, whenever possible, the same distribution of data for training, validation and testing used in the first iteration, and computed Stanza’s performance based on the *Unlabeled Attachment Score*¹³ (UAS) and *Labeled Attachment Score*¹⁴ (LAS). The UAS accuracy increased from 94.46% at the first run to 95.78% in the last (6th) iteration, becoming 1,32% better. For LAS, the final accuracy (6th run) achieved 94,62%, increasing 0,76% from the first run accuracy of 93,86%. The increase of the dependency relation measures indicates that the model’s ability to capture the syntactic structures of the tweets has improved as we incorporate news tweet into the training sets. For comparison purposes, the accuracy of the best model for journalist texts in Portuguese was also around 96% (UAS) and 95% (LAS) [Lopes and Pardo 2024]. Figure 3 depicts the overall distribution of the dependency relations (without subrelations) in DANTEStocks.

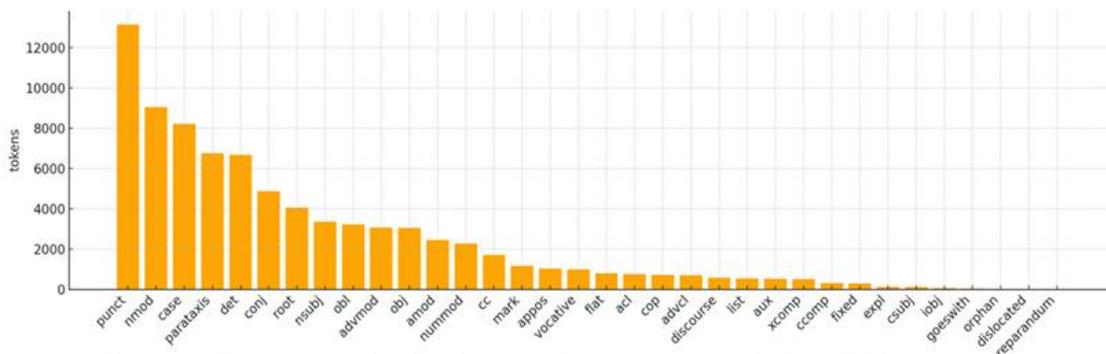


Figure 3. Frequency distribution of UD *deprel* tagset in DANTEStocks.

¹² The 8,418 sentences were split into training, development, and test sets, with 70% (5,893 sentences), 10% (842 sentences), and 20% (1,683 sentences) of the corpus, respectively.

¹³ UAS indicates the accuracy of the *head* ignoring the relation’s name (*deprel*) [Nivre and Fang, 2017].

¹⁴ LAS evaluates the output of a parser by considering how many words have been assigned both the correct syntactic *head* and the correct label, ignoring subrelations [Nivre and Fang 2017].

5. Reliability of Annotation

To provide a reliability measure of the annotation of DANTEStocks, a second NLP expert (also with UD-annotation experience) manually reviewed the automatic annotation of 100 random tweets based on the same guidelines [Duran 2022; Di Felippo *et al.* 2024]. The dependency-trees analyzed by the additional annotator could be from the reference subcorpus or generated by Stanza in one of its interactions. The *Inter-Annotator Agreement* (IAA) score was calculated by using the *Kappa* coefficient [Cohen, 1960; Carletta, 1996] in two different settings [Barbosa 2024]. In the first, the focus was to evaluate the annotation of *head* and *deprel* separately. The *Kappa* results for *head* and *deprel* were 0.96 and 0.97, respectively. In the second setting, the evaluation aimed at the combination of *head* and *deprel*, obtaining the *Kappa* score of 0.95. The IAA per *deprel* was measured by using the *total agreement* score [Sobrevilha Cabezudo 2015], since *Kappa* is not appropriate given the unbalanced distribution of the relations. We obtained the *total agreement* of 100% for more than half of the 46 different *deprels* (including subrelations) that occur in the sample of 100 tweets. Out of the 1.743 annotated relations, there are 42 cases of disagreement. The most frequent conflict was between *obl* and *nmod*. Some of them were caused by different but potential interpretations about the functional role of the prepositional phrase (in bold) in structure like “*arrisque vd em #petr4*” (“risk selling in #petr4”). While one annotator attached “petr4” to the verb via *obl*, functioning as a non-core (oblique) argument or adjunct, the other assumed that “petr4” is a modifier of the noun “*vd*” (“*venda*”), being attached to it by *nmod*. It is also interesting that, among the 22 *deprels* with *total agreement* different from 100%, 12 of them contain subrelations, indicating that the annotation is more complex when using language-specific relations. Even though a small-scale evaluation, the results indicate that the overall IAA was otherwise quite high, especially for the challenging task. This might be due to the large and detailed recommendations of our guidelines for the syntactic annotation of the tweets.

6. Final Remarks and Future Work

We described our effort on building the first BP treebank for Twitter microtext, annotated within the framework of UD. The contributions are the treebank itself, the instantiation of the UD guidelines for stock market tweets in BP, and the customization of a current state-of-the-art parser for tweets. Our main difficulty was interpreting the tweets, due to the medium- and domain-lexical phenomena and uncommon constructions. Thus, despite the constant help of domain experts, we can say that the dependency annotation of many tweets in DANTEStocks (especially those with fragmentation, *e.g.*, aborted text) represents potential syntactic analysis of the tweets. Currently, the two annotators involved in this work are analyzing the disagreements to assign a consensual *deprel* for each case and to make the treebank available soon. The guidelines for the syntactic UD-based annotation of DANTEStocks and the treebank itself (*beta version*) are available at the POeTiSA project webpage (<https://sites.google.com/icmc.usp.br/poetisa/>).

Acknowledgements. 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. The project was also supported by 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.

References

- Barbosa, B. K. S. (2024). Descrição sintático-semântica de nomes predicadores em tweets do mercado financeiro em português. Dissertação de Mestrado. Programa de Pós-graduação em Linguística, Universidade Federal de São Carlos, São Carlos/SP, 208p.
- Carletta, J. (1996). Assessing agreement on classification tasks: The kappa statistic. In *Computational Linguistics*, Volume 22, Number 2, pages 249–254. MIT Press.
- Cohen, J. (1960). A coefficient of agreement for nominal scales. In *Educational and Psychological Measurement*, Volume 20, Issue 1, pages 37-46.
- Di-Felippo, A.; Postali, C.; Ceregatto, G.; Gazana, L. S.; Roman, N. T. (2022). Diretrizes de anotação de PoS tags em tweets do mercado financeiro: orientações para anotação em língua portuguesa segundo a abordagem *Universal Dependencies*. *Relatório Técnico do ICMC 438*. Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo. São Carlos-SP, 24p.
- Di-Felippo, A., Nunes, M. G. V., Barbosa, B. K. S. (2024). Diretrizes de anotação de relações de dependência em tweets do mercado financeiro. *Relatório Técnico do ICMC 446*. Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo. São Carlos-SP, Abril, 70p.
- Duran, M.S. (2021). Manual de Anotação de PoS tags: orientações para anotação de etiquetas morfossintáticas em Língua Portuguesa, seguindo as diretrizes da abordagem *Universal Dependencies* (UD). *Relatório Técnico do ICMC 434*. ICMC, USP. São Carlos-SP, 55p.
- Duran, M.S. (2022). Manual de Anotação de Relações de Dependência - Versão Revisada e Estendida: Orientações para anotação de relações de dependência sintática em Língua Portuguesa, seguindo as diretrizes da abordagem *Universal Dependencies* (UD). *Relatório Técnico do ICMC 440*. ICMC, USP. São Carlos-SP, 166p.
- Duran, M. S., Lopes, L., Nunes, M.G.V., Pardo, T. A. S. (2023). The Dawn of the Porttinari Multigenre Treebank: Introducing its Journalistic Portion. In *Proceedings of the 14th Symposium in Information and Human Language Technology*, pages 115-124. Belo Horizonte/MG. SBC.
- Krumm, J., Davis, N. Narayanaswami, C. (2009). User-Generated Content. In *IEEE Pervasive Computing*, Volume 7, Issue 4, pages. 10 – 11, IEEE, 2009.
- Lopes, L., Duran, M. S.; Fernandes, P. H. L.; Pardo, T. A. S. (2022). PortiLexicon-UD: a Portuguese Lexical Resource according to Universal Dependencies Model. In *Proceedings of the 13th International Conference on Language Resources and Evaluation* (LREC), pages 6635 6643, Marseille, France. ELRA.
- Lopes, L.; Pardo, T. A. S. 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* (PROPOR), pages 401-410, Santiago de Compostela, Galiza. ACL.
- Luhn, H.P. (1957). A statistical approach to mechanized encoding and searching of literary information. In *IBM Journal of Research and Development*, Volume 1, Issue 4, pages 309-317. ISSN 0018-8646. doi:10.1147/rd.14.0309
- Macqueen, J. (1967) Some methods for classification and analysis of multivariate observations. In *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*. [S.l.], v. 1, n. 14, p. 281–297.

- Nivre, J., Fang, C.-T. (2017). Universal Dependency evaluation. In *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017)*, pages 86–95, Gothenburg, Sweden. ACL.
- Nivre, J., *et al.* (2016). Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC)*, pages 1659–1666, Portorož, Eslovênia. ELRA.
- Nivre, J. *et al.* (2020). Universal Dependencies v2: an evergrowing multilingual treebank collection. In *Proceedings of the 12th International Conference on Language Resources and Evaluation Conference (LREC)*, pages 4034-4043. Marseille, França. ELRA.
- Qi, P., Zhang, Y., Zhang, Y., Bolton, J., Manning, C. D. (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 (ACL) (System Demonstrations)*, pages 101-108. Online. ACL.
- Rademaker, A., Chalub, F., Real, L., Freitas, C., Bick, E., Paiva, V. de. (2017). Universal Dependencies for Portuguese. In *Proceedings of the 4th International Conference on Dependency Linguistics (Depling)*, pages 197–206, Pisa, Italy. Linköping University Electronic Press.
- Sanguinetti, M. *et al.* (2023). Treebanking user-generated content: a UD based overview of guidelines, corpora and unified recommendations. In *Lang Resources & Evaluation*, Volume. 57, Issue 2, pages 493–544. Springer-Verlag, Berlin, Heidelberg.
- Silva, E.H.; Pardo, T.A.S.; Roman, N.T.; Di Felippo, A. (2021). *Universal Dependencies for tweets in Brazilian Portuguese: tokenization and Part-of-Speech tagging*. In *Proceedings of the 18th National Meeting on Artificial and Computational Intelligence (ENIAC)*, pages. 434-445, Online. SBC.
- Scandarolli, C. L., Di-Felippo, A., Roman, N. T., Pardo, T. A. S. (2023). Tipologia de fenômenos ortográficos e lexicais em CGU: o caso dos tweets do mercado financeiro. In *Anais da VIII Jornada de Descrição do Português (JDP) (Evento integrante do XIV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana -STIL)*, p. 240-248, Belo Horizonte/MG, Brasil. SBC.
- Sobrevilla Cabezudo, M.A., Maziero, E.G., Souza, J.W.C., Dias, M.S., Cardoso, P.C.F., Balage Filho, P.P., Agostini, V., Nóbrega, F.A.A., Barros, C.D., Di Felippo, A., Pardo, T.A.S. (2015). Anotação de sentidos de verbos em textos jornalísticos do *corpus* CSTNews. In *Revista de Estudos da Linguagem (RELIN)*, Volume 23, Número 3, p. 797-832.
- Straka, M. (2018). UDPipe 2.0 prototype at CoNLL 2018 UD shared task. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 197–207. Brussels, Belgium. ACL.

Geração Automática de Perguntas em Português do Brasil Usando os Modelos PTT5 e FLAN-T5

Tiago Felipe V. Braga¹, Bruno Cardoso Coutinho², Hilário Tomaz Alves de Oliveira¹

¹Programa de Pós-graduação em Computação Aplicada (PPComp)
Instituto Federal do Espírito Santo (IFES) – Campus Serra

²Coordenadoria do Curso Técnico em Informática
Instituto Federal do Espírito Santo (IFES) – Campus Serra

tiagofvx@gmail.com, {bccout, hilario.oliveira}@ifes.edu.br

Abstract. *This paper performs a comparative analysis of the pre-trained neural models of PTT5 and FLAN-T5 for Brazilian Portuguese automatic question generation. To this end, two datasets, PIRÁ and FairyTaleQA, were used to evaluate the ability of these models to generate questions from two scenarios: (i) considering only the context and (ii) using the context and the expected answer. The ROUGE-L and BERTScore measures were used to assess the generated questions, in addition to an analysis based on GPT-4o. The results demonstrated that the PTT5_{Large} model consistently outperformed the other models, generating 93.06% of valid questions in PIRÁ and 82.32% in FairyTaleQA based on the GPT-4o evaluation.*

Resumo. *Este artigo apresenta uma análise comparativa dos modelos neurais pré-treinados do PTT5 e FLAN-T5 para a geração automática de perguntas em português do Brasil. Para isso, foram utilizados dois conjuntos de dados, PIRÁ e FairyTaleQA, para avaliar a capacidade desses modelos de gerar perguntas a partir de dois cenários: (i) considerando apenas o contexto e (ii) usando o contexto e a resposta esperada. As medidas do ROUGE-L e do BERTScore foram usadas para avaliar as perguntas geradas, além de uma análise baseada no GPT-4o. Os resultados demonstram que o modelo PTT5_{Large} apresentou consistentemente desempenho superior aos demais modelos, gerando 93,06% de perguntas válidas no PIRÁ e 82,32% no FairyTaleQA na avaliação baseada no GPT-4o.*

1. Introdução

A Geração de Perguntas (QG, do inglês *Question Generation*) é uma tarefa da área de Processamento de Linguagem Natural (PLN), que envolve a criação automática de perguntas a partir de um dado texto ou conjuntos de dados textuais [Zhang et al. 2021]. Usando técnicas de PLN e algoritmos de Aprendizado de Máquina (AM), os sistemas de QG visam gerar perguntas gramaticalmente corretas e contextualmente relevantes [da Rocha Junqueira et al. 2024]. Diante da grande abundância de informações digitais, sistemas de QG possuem diversas potenciais áreas de aplicação [Mulla and Gharpure 2023]. Na área da educação, a aplicação de abordagens de QG pode contribuir para o desenvolvimento de materiais de avaliação, questionários práticos, no desenvolvimento de sistemas de tutoria, aprimorando processos de aprendizagem e

avaliação [Kurdi et al. 2020]. No âmbito dos sistemas de perguntas e respostas (QA, do inglês Question Answering), abordagens de QG têm sido usadas para o treinamento de modelos com pouca supervisão ou para fins de aumento de dados [Puri et al. 2020].

Apesar do crescente interesse em pesquisas envolvendo a tarefa de QG, a maioria desses estudos concentra-se predominantemente na língua inglesa, onde há diversos recursos e bases de dados disponíveis para experimentação e desenvolvimento [Zhang et al. 2021, Mulla and Gharpure 2023]. Em contrapartida, as pesquisas focadas na língua portuguesa, especialmente para o português do Brasil, ainda são limitadas, resultando em uma escassez tanto de estudos quanto de bases de dados [da Rocha Junqueira et al. 2024, Leite et al. 2024]. Essa lacuna impõe desafios adicionais para o avanço no desenvolvimento e na aplicação prática de sistemas de QG, uma vez que a adaptação de modelos e técnicas desenvolvidas para o inglês nem sempre se traduzem diretamente em resultados eficazes em outros idiomas, dada a complexidade e as particularidades linguísticas inerentes de linguagem natural.

Este artigo tem como objetivo investigar a aplicação de modelos neurais de linguagem pré-treinados baseados na arquitetura *Transformers* [Vaswani 2017], mais especificamente os modelos *PTT5* [Carmo et al. 2020] e *FLAN-T5* [Chung et al. 2024] para a tarefa de QG em português do Brasil. Para isso, foi realizado o ajuste fino desses modelos em suas arquiteturas *small*, *base* e *large*, utilizando as bases de dados do PIRÁ [Paschoal et al. 2021], nativa em português, e uma versão traduzida da base de dados FairytaleQA [Leite et al. 2024] para o português do Brasil. Os experimentos foram realizados em dois cenários: no primeiro, as perguntas foram geradas a partir somente de um dado contexto; no segundo, as perguntas foram geradas considerando tanto o contexto quanto uma resposta prévia. A avaliação dos resultados foi realizada por meio das medidas automáticas do ROUGE-L e BERTScore, que são comumente utilizadas para avaliar abordagens de QG em termos de similaridade léxica e semântica das perguntas geradas com as perguntas de referência. Além disso, foi realizado um experimento adicional utilizando o modelo *GPT-4o* para avaliar as perguntas geradas. Esse experimento teve como objetivo complementar as avaliações quantitativas anteriores, proporcionando uma análise adicional da qualidade das perguntas geradas.

As principais contribuições deste artigo incluem: **(i)** o ajuste fino e a avaliação de diferentes arquiteturas dos modelos *PTT5* e *FLAN-T5* para a tarefa de QG em português do Brasil; e **(ii)** uma extensa investigação considerando duas bases de dados, PIRÁ e FairytaleQA, e duas variações da tarefa. O código-fonte desenvolvido neste trabalho está público em um repositório do GitHub¹.

2. Trabalhos Relacionados

As abordagens de geração de perguntas podem ser classificadas em métodos convencionais e baseados em modelos neurais [Zhang et al. 2021]. Os métodos convencionais de QG baseiam-se principalmente na aplicação de regras heurísticas para transformar os textos em perguntas relacionadas. Recentemente, com a evolução das arquiteturas de redes neurais profundas, houve uma mudança de paradigma na tarefa para a adoção de modelos neurais, permitindo assim, o desenvolvimento de abordagens orientadas a dados e completamente treináveis, na qual a seleção de conteúdo e a construção de perguntas podem

¹https://github.com/laicsiifes/question_generation_ptbr

ser otimizadas de forma combinada. Embora exista uma vasta literatura sobre QG em diversos idiomas [Kurdi et al. 2020, Zhang et al. 2021, Mulla and Gharpure 2023], por limitação de espaço, esta seção foca em trabalhos envolvendo o português do Brasil.

Em [Leite and Lopes Cardoso 2022], os autores apresentam um estudo que envolveu o treinamento do modelo *PTT5* para a geração de perguntas utilizando uma versão em português do conjunto de dados SQuAD 1.1. Os resultados obtidos foram encorajadores, com desempenho equiparável com a implementação em inglês do modelo *T5*, evidenciando a eficácia dos modelos baseados na arquitetura *Transformers* e estabelecendo *baselines* para futuras comparações para a tarefa de QG em português. Oliveira et al. [Oliveira et al. 2023] abordam o desafio de gerar e classificar distratores (opções incorretas) para questões de múltipla escolha em português. Os autores desenvolveram e combinam vários métodos de geração de distratores, incluindo extração baseada em contexto, manipulação numérica e similaridade semântica a partir de recursos como WordNet.

Junqueira et al. [da Rocha Junqueira et al. 2024] apresentaram uma investigação do desempenho dos modelos *T5*, *FLAN-T5* e *BART-PT* para a geração de perguntas factuais em português do Brasil. Para mitigar o problema da escassez de dados, foi utilizada uma versão em português brasileiro do SQuAD v1.1, obtida por meio de tradução automática. Leite et al. [Leite et al. 2024] realizaram a construção de versões traduzidas automaticamente da base de dados FairytaleQA, que é um conjunto de dados comumente usado para o desenvolvimento de sistemas de perguntas e respostas em inglês. Foram desenvolvidas versões do FairytaleQA para o português de Portugal, português do Brasil, espanhol e francês, que podem ser usadas em pesquisas da área de QG e QA. Além disso, foram realizados experimentos usando modelos neurais baseados na arquitetura *T5*.

Este trabalho difere dos anteriores ao: **(i)** treinar e avaliar diferentes tamanhos de arquitetura dos modelos *PTT5* e *FLAN-T5*, **(ii)** considerar dois cenários da tarefa de QG, **(iii)** adotar uma base de dados escrita nativamente em português (PIRÁ) e outra obtida por meio de tradução automática (FairytaleQA), e **(iv)** analisar o desempenho dos modelos usando uma abordagem com o modelo *GPT-4o*, além de tradicionais medidas de avaliação consideradas em trabalhos anteriores.

3. Materiais e Métodos

3.1. Bases de Dados

Neste trabalho foram utilizados dois conjuntos de dados, o PIRÁ [Paschoal et al. 2021] e o FairyTaleQA [Xu et al. 2022]. Essas bases de dados foram selecionadas por serem usadas em trabalhos da literatura na tarefa de geração de perguntas ou de sistemas de perguntas e respostas em português do Brasil. Além disso, elas possuem três componentes essenciais para a tarefa de QG: **(i)** contexto textual, **(ii)** pergunta associada e **(iii)** resposta correspondente.

O PIRÁ é uma base de dados bilíngue (português-inglês) focada em questões oceânicas e da costa brasileira. A base contém 2.261 textos extraídos de trechos de relatórios das Nações Unidas sobre o oceano e de resumos relacionados ao litoral brasileiro [Paschoal et al. 2021]. As perguntas e respostas foram criadas manualmente em um processo de revisão em pares por avaliadores humanos. Após uma análise da base de dados, foi observado que alguns exemplos não apresentam as respostas para as perguntas. Por

isso, esses exemplos foram removidos, já que a resposta é um elemento importante para os experimentos realizados neste trabalho.

O **FairyTaleQA** é uma base de dados comumente usada para avaliar sistemas de perguntas e respostas em inglês. Essa base foi criada por especialistas em educação e é composta por textos narrativos infantis. Leite et al. [Leite et al. 2024] realizaram um processo de tradução do FairyTaleQA para diversos idiomas, incluindo o português de Portugal e do Brasil. Neste trabalho, foi utilizada a versão traduzida para o português do Brasil, que compreende 10.580 perguntas e respostas derivadas de 278 histórias infantis.

Na Tabela 1 são apresentadas para cada base de dados as estatísticas do total de exemplos em cada conjunto (treinamento, validação e teste) e o tamanho médio e desvio padrão do total de palavras para cada componente (contexto, pergunta e resposta). Para gerar essas estatísticas, foi utilizada a ferramenta spaCy² para o processamento dos textos.

Tabela 1. Estatística das bases de dados do PIRÁ e FairyTaleQA

Base de Dados	Conjunto	Exemplos	Componente	Média de Palavras (Desvio Padrão)
PIRÁ	Treino	1.756	Contexto	274,73 (141,41)
			Pergunta	13,83 (5,62)
			Resposta	14,32 (11,76)
	Validação	215	Contexto	273,98 (157,08)
			Pergunta	13,65 (5,38)
			Resposta	15,04 (12,06)
	Teste	216	Contexto	250,58 (128,98)
			Pergunta	13,36 (5,68)
			Resposta	14,92 (14,50)
FairyTaleQA	Treino	8.548	Contexto	182,51 (94,53)
			Pergunta	10,23 (3,38)
			Resposta	6,98 (5,73)
	Validação	1.025	Contexto	170,08 (74,18)
			Pergunta	10,93 (3,40)
			Resposta	7,52 (5,96)
	Teste	1.007	Contexto	168,92 (73,77)
			Pergunta	10,48 (3,30)
			Resposta	6,80 (5,31)

3.2. Modelos Avaliados

Neste trabalho, foram avaliados os modelos *PTT5* e o *FLAN-T5*, baseados na arquitetura *Text-to-Text Transfer Transformer (T5)* [Raffel et al. 2020]. Apesar de existirem diferentes tamanhos de arquitetura, as três comumente usadas são “*small*”, “*base*” e “*large*”. Elas possuem um número crescente de parâmetros, o que geralmente resulta em maior capacidade de aprendizado, mas também em um maior custo computacional. Esses modelos foram escolhidos devido ao seu desempenho promissor em tarefas de PLN e por terem sido explorados em trabalhos anteriores.

O *PTT5* é uma adaptação do modelo *T5*, especificamente pré-treinada para o português do Brasil [Carmo et al. 2020]. O modelo foi pré-treinado no corpus BrWac [Wagner Filho et al. 2018], uma extensa coleção de páginas *web* em português, contendo aproximadamente 2,7 bilhões de *tokens*. Foram utilizados os modelos *PTT5_{Small}*, que possui aproximadamente 60 milhões de parâmetros, o *PTT5_{Base}*, com cerca de 220 milhões de parâmetros, e o *PTT5_{Large}*, que apresenta aproximadamente 740 milhões de parâmetros.

²<https://spacy.io/>

O *FLAN-T5* [Chung et al. 2024] é uma versão aprimorada do *T5* pré-treinado em múltiplas tarefas de PLN. Esse modelo foi pré-treinado majoritariamente em documentos em inglês, mas possui suporte a outros idiomas, como o português. Foram avaliadas três variantes deste modelo: o *FLAN-T5_{Small}* com cerca de 80 milhões de parâmetros, o *FLAN-T5_{Base}* contendo aproximadamente 250 milhões de parâmetros, e o *FLAN-T5_{Large}* apresentando cerca de 780 milhões de parâmetros. Sua inclusão tem o objetivo de avaliar como um modelo com treinamento diversificado se comporta em comparação a modelos especializados em um único idioma e tarefa.

3.3. Metodologia Experimental

A metodologia experimental utilizada neste trabalho envolveu o desenvolvimento, ajuste fino e a avaliação dos modelos investigados, especificamente ajustados para dois cenários da tarefa de geração de perguntas, conforme ilustrado na Figura 1. Para cada cenário, seis modelos foram treinados, considerando os três tamanhos de arquitetura e os dois modelos *PTT5* e *FLAN-T5*. No primeiro cenário, foi analisada a variação da tarefa que gera perguntas a partir somente do contexto, como entrada. Já no segundo cenário, os modelos recebem tanto o contexto quanto uma resposta como entrada e devem gerar uma pergunta como saída.

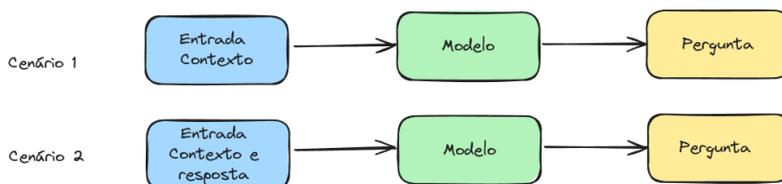


Figura 1. Cenários da tarefa de QG analisados.

Os modelos *PTT5* e *FLAN-T5* foram implementados usando a biblioteca *Transformers*³. O tamanho de entrada máximo foi definido para 512 *tokens*, enquanto a saída foi configurada para no máximo 40 *tokens*. Durante o treinamento, os modelos foram ajustados por no máximo 20 épocas, sendo utilizada a estratégia de parada antecipada com uma paciência de 5 épocas. Para mitigar o sobreajuste dos modelos, ao final de cada época, o modelo treinado é aplicado no conjunto de validação e é computada a medida de ROUGE-L, sendo salvo somente o modelo com maior valor. Durante a geração das perguntas, foi utilizado o algoritmo de *Beam Search* com uma largura de tamanho 5. Esses valores foram definidos a partir da análise de trabalhos anteriores.

A avaliação do desempenho dos modelos foi realizada por meio de duas abordagens: a aplicação de métricas automáticas de similaridade e uma avaliação com base no modelo de linguagem *GPT-4o*. Para a avaliação automática, foi utilizada a métrica Recall-Oriented Understudy for Gisting Evaluation Longest Common Subsequence (ROUGE-L) [Lin 2004], que mensura a similaridade com base na maior sequência de palavras em comum entre as perguntas geradas e as perguntas de referência. Adicionalmente, foi utilizada a métrica BERTScore [Zhang et al. 2019], que calcula a similaridade de cosseno a partir das representações em embeddings extraídas do modelo *Bidirectional Encoder Representations for Transformers* (BERT).

³<https://huggingface.co/docs/transformers/index>

Para uma avaliação mais holística e contextualmente relevante, foi realizada uma análise usando o modelo *GPT-4o*. Essa avaliação foi pensada porque, dado um contexto específico, é possível gerar múltiplas perguntas válidas que não necessariamente precisam ser idênticas à pergunta de referência presente nas bases de dados usadas nos experimentos. Esta situação é particularmente relevante no Cenário 1, onde apenas o contexto é fornecido como entrada para o modelo. Neste caso, diversas perguntas podem ser consideradas válidas, desde que sejam respondíveis com base no contexto fornecido. Em contraste, no Cenário 2, onde o contexto e a resposta esperada são fornecidos como entrada, a pergunta gerada deve ser semanticamente equivalente à pergunta de referência.

O processo de avaliação utilizando o *GPT-4o* foi inspirado na técnica de *Retrieval Augmented Generation* (RAG). Esta técnica consiste em fornecer um contexto e uma pergunta para um modelo de linguagem de grande escala (LLM, do inglês *Large Language Model*), solicitando que ele responda à pergunta usando apenas o contexto fornecido ou sinalize caso não seja possível [Chen et al. 2024]. Seguindo esta abordagem, foi criado um *prompt*⁴ contendo o contexto original e a pergunta gerada pelos modelos avaliados. Este *prompt* foi então submetido ao *GPT-4o*, com a instrução de responder à pergunta utilizando somente o contexto fornecido ou indicar a impossibilidade de resposta. Com base nas respostas do *GPT-4o*, foi calculado o percentual de perguntas válidas (aquelas que o LLM conseguiu responder com base no contexto) e inválidas (as que não puderam ser respondidas) para cada modelo avaliado. Deste modo, foi possível analisar se as perguntas geradas foram relevantes ao contexto, ainda que diferentes da pergunta de referência. Embora esta análise seja automatizada, foi realizada uma inspeção manual em amostras das saídas do *GPT-4o* para verificar sua confiabilidade. Foi observado que, em geral, o LLM identificava corretamente as perguntas válidas e inválidas.

4. Resultados

Na Tabela 2 são apresentados os resultados dos experimentos, considerando os cenários 1 e 2, com base nas medidas de avaliação do ROUGE-L e BERTScore. No Cenário 1 (apenas contexto), o *PTT5_{Base}* e o *PTT5_{Large}* apresentaram os melhores desempenhos para as bases de dados do FairyTaleQA e PIRA, respectivamente. No Cenário 2 (contexto e resposta esperada), o *PTT5_{Large}* superou os demais modelos em ambas as bases. Fica evidente que os modelos *PTT5* consistentemente obtiveram melhores resultados do que os modelos *FLAN-T5* em ambos os conjuntos de dados e cenários, sugerindo que o pré-treinamento específico em português confere vantagens na tarefa de geração de perguntas.

Comparando os resultados obtidos em ambos os cenários de avaliação, observa-se que os modelos apresentaram melhores desempenhos no Cenário 2 em comparação com o Cenário 1. Isso acontece porque no Cenário 2, como a resposta é dada como entrada, ela guia os modelos a gerarem perguntas para aquele contexto e resposta. Assim, a pergunta gerada precisa ser semanticamente equivalente à pergunta de referência. Tal situação não ocorre no Cenário 1, já que é somente dado o contexto como entrada e, para um mesmo contexto, é possível gerar diversas perguntas válidas. Por isso, para melhor avaliar o Cenário 1, foram realizadas as análises usando o modelo *GPT-4o*.

Na Tabela 3 são apresentados os resultados da avaliação dos modelos no Cenário 1 usando o *GPT-4o*. Os resultados obtidos apresentam um padrão similar ao primeiro ex-

⁴O *prompt* usado está disponível no repositório do projeto.

Tabela 2. Resultados dos experimentos nos cenários 1 e 2.

Cenário	Base de Dados	Modelo	ROUGE-L	BERTScore
1	FairyTaleQA	<i>PTT5_{Small}</i>	0,2491	0,3976
		<i>PTT5_{Base}</i>	0,2699	0,4137
		<i>PTT5_{Large}</i>	0,2668	0,4093
		<i>FLAN-T5_{Small}</i>	0,2412	0,3469
		<i>FLAN-T5_{Base}</i>	0,2497	0,3590
		<i>FLAN-T5_{Large}</i>	0,2354	0,3448
	PIRÁ	<i>PTT5_{Small}</i>	0,2109	0,2982
		<i>PTT5_{Base}</i>	0,2266	0,3265
		<i>PTT5_{Large}</i>	0,2280	0,3449
		<i>FLAN-T5_{Small}</i>	0,1581	0,2099
		<i>FLAN-T5_{Base}</i>	0,1723	0,2404
		<i>FLAN-T5_{Large}</i>	0,2219	0,2988
2	FairyTaleQA	<i>PTT5_{Small}</i>	0,4230	0,5429
		<i>PTT5_{Base}</i>	0,4786	0,5906
		<i>PTT5_{Large}</i>	0,4938	0,6057
		<i>FLAN-T5_{Small}</i>	0,3190	0,4203
		<i>FLAN-T5_{Base}</i>	0,3672	0,4611
		<i>FLAN-T5_{Large}</i>	0,3810	0,4884
	PIRÁ	<i>PTT5_{Small}</i>	0,2625	0,3635
		<i>PTT5_{Base}</i>	0,3506	0,4505
		<i>PTT5_{Large}</i>	0,3640	0,4656
		<i>FLAN-T5_{Small}</i>	0,1680	0,2220
		<i>FLAN-T5_{Base}</i>	0,1934	0,2579
		<i>FLAN-T5_{Large}</i>	0,2620	0,3257

perimento, mas com algumas diferenças importantes. O *PTT5_{Large}* obteve o melhor desempenho em ambas as bases de dados, com 82,32% das perguntas geradas sendo consideradas válidas no FairyTaleQA e 93,06% no PIRÁ. É possível observar uma divergência entre as medidas automáticas e a avaliação *GPT-4o*, particularmente no PIRÁ. Enquanto as medidas do ROUGE-L e BERTScore indicaram valores menores para o PIRÁ em comparação com o FairyTaleQA, a avaliação *GPT-4o* mostrou uma tendência oposta, com percentuais mais altos de perguntas válidas no PIRÁ.

Tabela 3. Resultados da avaliação do Cenário 1 usando o *GPT-4o*.

Base de Dados	Modelo	Válida	Inválida	% Válida
FairyTaleQA	<i>PTT5_{Small}</i>	680	327	67,53
	<i>PTT5_{Base}</i>	726	281	72,10
	<i>PTT5_{Large}</i>	829	178	82,32
	<i>FLAN-T5_{Small}</i>	451	556	44,79
	<i>FLAN-T5_{Base}</i>	525	482	52,14
	<i>FLAN-T5_{Large}</i>	719	288	71,40
PIRÁ	<i>PTT5_{Small}</i>	135	81	62,50
	<i>PTT5_{Base}</i>	199	17	92,13
	<i>PTT5_{Large}</i>	201	15	93,06
	<i>FLAN-T5_{Small}</i>	122	94	56,48
	<i>FLAN-T5_{Base}</i>	138	78	63,89
	<i>FLAN-T5_{Large}</i>	197	19	91,20

Na Figura 2 é apresentado um exemplo extraído da base de dados do PIRÁ, contendo o contexto, as perguntas geradas pelos modelos *PTT5* e a saída da análise usando o *GPT-4o*. Nesse exemplo, é possível ver que os modelos *PTT5_{Large}* e *PTT5_{Base}* foram capazes de gerar perguntas que podem ser respondidas pelo contexto, sendo assim consideradas *válidas*. Por outro lado, o modelo *PTT5_{Small}* gerou uma pergunta confusa sobre a Petróleo Brasileiro SA não ter comentado sobre a estimativa de produção revisada. Ape-

sar de ser mencionado no contexto, não está explícito nele o porquê disso. Sendo assim, considerada *inválida* pela avaliação do *GPT-4o*.

Contexto: O BG GROUP produziu recentemente uma nova estimativa na descoberta de óleo de Tupi na Bacia de Santos, afirmando que o campo contém 12-30 bilhões boe ou mais. Por um lado, a Petroleo Brasileiro SA de Petróleo (Petrobras) não comentou sobre a estimativa de produção revisada. Pode-se lembrar que a Petrobras relatou a descoberta para ser de 8 bilhões bbl de luz em bruto em 2007. Enquanto isso, alegações de BG foram produzidas em uma declaração sobre a estratégia de crescimento a longo prazo da empresa, lançada à frente da apresentação de seus resultados do quarto trimestre.

PTT5_{Small}: Por que a Petroleo Brasileiro SA de Petróleo (Petrobras) não comentou sobre a estimativa de produção revisada?

Avaliação: INVÁLIDA.

PTT5_{Base}: Qual foi a estimativa do BG Group para a descoberta de óleo de Tupi na Bacia de Santos?

Avaliação: VÁLIDA.

PTT5_{Large}: Qual a estimativa do BG Group para a quantidade de óleo de Tupi na Bacia de Santos?

Avaliação: VÁLIDA.

Figura 2. Exemplo de contexto extraído do PIRÁ e perguntas consideradas válidas e inválidas pelo *GPT-4o*.

5. Considerações Finais e Trabalhos Futuros

Neste trabalho, foi realizada uma análise comparativa dos modelos *PTT5* e *FLAN-T5* para a tarefa de geração automática de perguntas. Para isso, foram utilizadas as bases de dados do PIRÁ e uma versão traduzida do FairyTaleQA para o português do Brasil. O desempenho dos modelos foi avaliado usando uma abordagem tradicional, considerando as medidas de avaliação do ROUGE-L e do BERTScore. Além dessa abordagem, foi realizada uma análise das perguntas geradas pelos modelos usando o *GPT-4o*, avaliando se as perguntas geradas poderiam ser respondidas somente a partir do contexto fornecido. Os resultados experimentais demonstraram que o modelo *PTT5_{Large}* obteve os melhores resultados em quase todos os cenários avaliados. Os resultados obtidos indicam a eficácia do pré-treinamento específico em português, evidenciada pelo desempenho superior consistente dos modelos *PTT5* em comparação com os modelos *FLAN-T5*.

Apesar dos resultados encorajadores obtidos, o trabalho apresenta diversas limitações, que serão melhor exploradas. Dentre elas, pode-se destacar duas linhas de pesquisa futuras: (i) investigar o desempenho de LLMs, como o Llama 3 [Touvron et al. 2023], Gemma [Team et al. 2024] e Sabiá [Almeida et al. 2024]; e (ii) realizar uma avaliação humana para complementar as avaliações automáticas realizadas.

Agradecimentos

Os autores agradecem ao Ifes, apoio da FAPES e CAPES (processo 2021-2S6CD, nº FAPES 132/2021) por meio do PDPG (Programa de Desenvolvimento da Pós-Graduação, Parcerias Estratégicas nos Estados).

Referências

- Almeida, T. S., Abonizio, H., Nogueira, R., and Pires, R. (2024). Sabiá-2: A new generation of portuguese large language models. *arXiv preprint arXiv:2403.09887*.
- Carmo, D., Piau, M., Campiotti, I., Nogueira, R., and Lotufo, R. (2020). Ptt5: Pre-training and validating the t5 model on brazilian portuguese data. *arXiv preprint arXiv:2008.09144*.
- Chen, J., Lin, H., Han, X., and Sun, L. (2024). Benchmarking large language models in retrieval-augmented generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17754–17762.
- Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, Y., Wang, X., Dehghani, M., Brahma, S., et al. (2024). Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.
- da Rocha Junqueira, J., Corrêa, U. B., and Freitas, L. (2024). Transformer models for brazilian portuguese question generation: An experimental study. In *The International FLAIRS Conference Proceedings*, volume 37.
- Kurdi, G., Leo, J., Parsia, B., Sattler, U., and Al-Emari, S. (2020). A systematic review of automatic question generation for educational purposes. *International Journal of Artificial Intelligence in Education*, 30:121–204.
- Leite, B. and Lopes Cardoso, H. (2022). Neural question generation for the portuguese language: A preliminary study. In *EPIA Conference on Artificial Intelligence*, pages 780–793. Springer.
- Leite, B., Osório, T. F., and Cardoso, H. L. (2024). Fairytaleqa translated: Enabling educational question and answer generation in less-resourced languages. *arXiv preprint arXiv:2406.04233*.
- Lin, C.-Y. (2004). ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.
- Mulla, N. and Gharpure, P. (2023). Automatic question generation: a review of methodologies, datasets, evaluation metrics, and applications. *Progress in Artificial Intelligence*, 12(1):1–32.
- Oliveira, H. G., Caetano, I., Matos, R., and Amaro, H. (2023). Generating and ranking distractors for multiple-choice questions in portuguese. In *SLATE*, pages 4–1.
- Paschoal, A. F., Pirozelli, P., Freire, V., Delgado, K. V., Peres, S. M., José, M. M., Nakasato, F., Oliveira, A. S., Brandão, A. A., Costa, A. H., et al. (2021). Pirá: A bilingual portuguese-english dataset for question-answering about the ocean. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 4544–4553.
- Puri, R., Spring, R., Shoeybi, M., Patwary, M., and Catanzaro, B. (2020). Training question answering models from synthetic data. In Webber, B., Cohn, T., He, Y., and Liu, Y., editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5811–5826, Online. Association for Computational Linguistics.

- Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.
- Team, G., Mesnard, T., Hardin, C., Dadashi, R., Bhupatiraju, S., Pathak, S., Sifre, L., Rivi re, M., Kale, M. S., Love, J., et al. (2024). Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.
- Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozi re, B., Goyal, N., Hambro, E., Azhar, F., et al. (2023). Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.
- Vaswani, A. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*.
- Wagner Filho, J. A., Wilkens, R., Idiart, M., and Villavicencio, A. (2018). The brwac corpus: A new open resource for brazilian portuguese. In *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*.
- Xu, Y., Wang, D., Yu, M., Ritchie, D., Yao, B., Wu, T., Zhang, Z., Li, T., Bradford, N., Sun, B., Hoang, T., Sang, Y., Hou, Y., Ma, X., Yang, D., Peng, N., Yu, Z., and Warschauer, M. (2022). Fantastic questions and where to find them: FairytaleQA – an authentic dataset for narrative comprehension. In Muresan, S., Nakov, P., and Villavicencio, A., editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 447–460, Dublin, Ireland. Association for Computational Linguistics.
- Zhang, R., Guo, J., Chen, L., Fan, Y., and Cheng, X. (2021). A review on question generation from natural language text. *ACM Trans. Inf. Syst.*, 40(1).
- Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., and Artzi, Y. (2019). Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Sumarização Automática de Artigos de Notícias em Português: Da Extração à Abstração com Abordagens Clássicas e Modelos de Neurais

Marcio Alves Sarmiento¹, Hilário Tomaz Alves de Oliveira¹

¹Programa de Pós-graduação em Computação Aplicada (PPComp)
Instituto Federal do Espírito Santo (IFES) – Campus Serra

malves.sarmiento@gmail.com, hilario.oliveira@ifes.edu.br

Abstract. *Automatic text summarization aims to generate a summary that captures the most relevant information from one or more textual documents. Although there have been significant advances in this field, research on documents written in Brazilian Portuguese remains limited. This article provides an analysis of various summarization approaches, ranging from classical baselines to extractive methods, including the fine-tuning of different architectures of the PPT5 and FLAN-T5 and the use of large language models for abstractive summarization. Experiments were conducted using three datasets of news articles written in Portuguese. The results showed that the models fine-tuned for the abstractive summarization achieved competitive performance based on ROUGE-L and BERTScore when compared to larger models like GPT-4o.*

Resumo. *A sumarização automática de texto tem como objetivo a criação de um resumo com as informações mais relevantes extraídas de um ou mais documentos textuais. Apesar dos avanços obtidos na área, pesquisas envolvendo documentos escritos em português do Brasil ainda são escassas. Este artigo apresenta uma análise envolvendo diferentes abordagens de sumarização, desde baselines clássicas, passando por sistemas extrativos, o ajuste fino de diferentes arquiteturas dos modelos PPT5 e FLAN-T5, até o uso de modelos de linguagem de larga escala para sumarização abstrativa. Experimentos foram realizados considerando três bases de dados de artigos de notícias escritas em português. Os resultados demonstraram que os modelos ajustados para a tarefa de sumarização abstrativa obtiveram resultados competitivos com base nas medidas do ROUGE-L e do BERTScore com modelos maiores, como o GPT-4o.*

1. Introdução

A crescente demanda por informações impulsiona o desenvolvimento de tecnologias capazes de processar e sintetizar grandes volumes de dados de forma rápida e eficiente. Em um cenário em que a produção de conteúdo digital online é cada vez mais abundante, torna-se cada vez mais desafiador para os leitores acompanhar todas as notícias relevantes [Levitin 2014, Zhang et al. 2024]. Nesse contexto, sistemas de Sumarização Automática de Texto (SAT) podem ser ferramentas úteis para auxiliar os usuários, oferecendo a capacidade de gerar resumos concisos que capturam as informações mais relevantes de um texto ou de múltiplos documentos relacionados, permitindo uma assimilação mais rápida do conteúdo [Lin and Ng 2019, Zhang et al. 2022].

A SAT é uma área de pesquisa em Processamento de Linguagem Natural (PLN) que busca gerar resumos de documentos textuais de forma automática. Existem duas abordagens principais para a tarefa de SAT: a *Extrativa* e a *Abstrativa* [Nenkova and McKeown 2012]. A *sumarização extrativa* seleciona as frases mais relevantes diretamente do texto original para compor o resumo, enquanto a *sumarização abstrativa* envolve a reescrita do conteúdo de forma mais condensada e frequentemente utiliza técnicas de geração de linguagem natural, sendo capaz de criar novas frases que não necessariamente aparecem no texto original [Zhang et al. 2022]. A sumarização pode ser aplicada tanto a um único documento (*monodocumento*) quanto a um conjunto de documentos (*multidocumento*).

Nos últimos anos, houve uma mudança no foco das pesquisas na área de SAT das abordagens extrativas para as abstrativas [Lin and Ng 2019]. Essa mudança foi impulsionada pelo desenvolvimento de algoritmos baseados em redes neurais profundas, especialmente na arquitetura *Transformer* [Vaswani 2017], capazes de geração de linguagem natural. Assim, diversas abordagens surgiram usando modelos neurais pré-treinados e, mais recentemente, modelos de linguagem de larga escala, do inglês *Large Language Models* (LLMs) [Zhang et al. 2022]. Contudo, apesar dos resultados promissores usando essas abordagens neurais, elas impõem diversos desafios, como a necessidade de grandes bases de dados para treinamento e demandam muitos recursos computacionais. Apesar dos avanços na área, a maioria das pesquisas tem como foco a língua inglesa, e poucos estudos têm sido dedicados ao português, especialmente para a sumarização abstrativa [Zhang et al. 2022]. Essa lacuna limita a aplicabilidade de sistemas de sumarização automática projetados especificamente para o português, que enfrenta a ausência de modelos e bases de dados para suportar essas pesquisas [Paiola et al. 2024].

Este artigo busca contribuir para o desenvolvimento da tarefa de SAT em português do Brasil por meio de uma investigação de algoritmos de sumarização aplicados a artigos de notícias. O estudo abrange desde o uso de sistemas extrativos, comumente usados como *baselines* de comparação, o ajuste fino dos modelos *PTT5* e *FLAN-T5* para sumarização abstrativa, até o uso de LLMs de código aberto e proprietários, como o GPT4-o¹, Llama 3 [Touvron et al. 2023] e o Gemma [Team et al. 2024]. Para isso, foram realizados experimentos em três bases de dados, o Temário e Recognasum para a tarefa de sumarização monodocumento e o CSTNews para sumarização multidocumento. O desempenho dos modelos foi avaliado usando as medidas de avaliação automáticas do ROUGE-L e BERTScore, que são usualmente adotadas na literatura.

Os códigos desenvolvidos e os resumos gerados neste trabalho estão públicos em https://github.com/laicsiifes/benchmark_ptbr_summ.

2. Trabalhos Relacionados

A literatura da área de SAT é vasta e existem diversos *surveys* que fornecem uma visão ampla do desenvolvimento da área desde a sua origem [Nenkova and McKeown 2012, Lin and Ng 2019, Zhang et al. 2022]. Por limitações de espaço, esta seção foca apenas em trabalhos que envolveram documentos escritos em português ou que usaram técnicas adotadas nos experimentos realizados neste estudo.

¹<https://openai.com/index/hello-gpt-4o/>

Diversos indicadores de relevância vêm sendo explorados para a execução da tarefa de sumarização extrativa [Leite and Rino 2008, Oliveira et al. 2016a]. Em sua maioria, esses indicadores baseiam-se em técnicas estatísticas, como frequência e centralidade, ou em heurísticas, como a posição das sentenças nos documentos. O estudo conduzido por Oliveira et al. [Oliveira et al. 2016a] avaliou diferentes técnicas para mensurar a relevância de sentenças em tarefas de SAT de artigos jornalísticos em inglês. Os autores analisaram os métodos individualmente e em combinação, utilizando-os como atributos em algoritmos de classificação. Leite e Rino [Leite and Rino 2008] investigaram uma abordagem combinando múltiplas *features* e algoritmos de aprendizado de máquina para a sumarização extrativa de documentos em português.

No trabalho de Sodr e e Oliveira [Sodr e and de Oliveira 2019], os autores investigaram a estrat egia de combinar alguns dos indicadores analisados por Oliveira et al. [Oliveira et al. 2016a] e aplicaram algoritmos de regress ao para estimar um escore de relev ancia das sentenas na tarefa de sumarizaao de artigos jornalísticos em português. Gomes e Oliveira [Gomes and de Oliveira 2019] propuseram um sistema usando Programação Linear Inteira (PLI) para sumarizaao extrativa multidocumento. O sistema desenvolvido usa bigramas como conceitos e aplica m etodos estatísticos tradicionais para identificar as informaoes mais relevantes para a construao do resumo.

Diferentemente dos trabalhos anteriores, Paiola et al. [Paiola et al. 2022] investigaram a tarefa de sumarizaao abstrativa. Os autores usaram diversas bases de dados em português (TeM ario, CSTNews, WikiLingua e XL-Sum) e um sistema de traduao para aplicar modelos treinados em ingl es. Em [Paiola et al. 2024], os autores apresentam a base de dados do RecognaSumm, um conjunto de dados contendo mais de 135 mil artigos de not cias para a tarefa de SAT. Os autores realizaram diferentes an lises da base de dados proposta e avaliaram o desempenho do modelo base do *PTT5* para estabelecer um desempenho de refer ncia para comparaoes futuras.

Este trabalho busca expandir os anteriores ao realizar uma an lise mais ampla considerando tr s bases de dados (TeM ario, RecognaSumm e CSTNews) para sumarizaao monodocumento e multidocumento, al m de envolver desde t cnicas de sumarizaao extrativas tradicionais de ponderaao das frases assim como os usados nos trabalhos em [Oliveira et al. 2016a, Sodr e and de Oliveira 2019], adaptaao do sistema de PLI proposto em [Gomes and de Oliveira 2019], ajuste fino e avaliaao de diferentes tamanhos de arquiteturas (*small*, *base* e *large*) dos modelos *PTT5* e *FLAN-T5* e o uso de LLMs de c digo aberto (*Llama3* e *Gemma2*) e propriet rios (*GPT-3.5* e *GPT-4o*).

3. Materiais e M etodos

3.1. Bases de Dados

Neste trabalho, foram utilizadas tr s bases de dados comumente usadas na literatura para a tarefa de SAT no dom nio de artigos de not cias escritas em português.

TeM ario. Esse conjunto de dados   formado por 100 textos jornalísticos, provenientes da Folha de S.Paulo e do Jornal do Brasil. Os artigos, que abordam uma variedade de temas, foram selecionados por sua linguagem clara e objetiva. Todos os textos possuem resumos elaborados por um especialista, o que garante a qualidade dos resumos de refer ncia [Pardo and Rino 2003].

CSTNews. Essa base de dados é formada por 50 conjuntos de notícias, cada um com aproximadamente quatro artigos sobre o mesmo tema, coletados manualmente em sites de notícias como Folha de São Paulo e Estadão. Essa abordagem permitiu a seleção de notícias com linguagem clara e acessível, provenientes de diferentes fontes sobre um mesmo assunto [Cardoso et al. 2011].

RecognaSumm. Com o objetivo de construir um conjunto de dados robusto para estudos de sumarização de textos, Paiola et al. [Paiola et al. 2024] coletaram 135.272 artigos de notícias usando sistemas de *web crawlers* personalizados. A diversidade temática dos artigos foi garantida pela coleta de dados em diferentes portais de notícias e categorias. A base de dados é dividida em três subconjuntos: treinamento, validação e teste. Por conta de limitações de *hardware*, foi feita uma filtragem no conjunto de treinamento, sendo removidos os artigos com resumos contendo menos do que 25 palavras.

Na Tabela 1 são apresentadas algumas estatísticas das bases de dados usadas nos experimentos. Para cada base, foi computado o total de documentos ou grupos, média e Desvio Padrão (DP) de frases e palavras no texto dos artigos.

Tabela 1. Estatísticas das bases de dados usadas nos experimentos.

Base de Dados	Conjunto	Docs / Grupos	Média (DP) Frases	Média (DP) Palavras
TeMário	Único	100	32,4 (10,38)	618,67 (163,93)
CSTNews	Único	50	47,06 (19,47)	939,56 (331,42)
RecognaSumm	Treino	64.347	27,07 (24,82)	527,33 (468,38)
	Validação	21.538	26,73 (24,15)	519,91 (458,68)
	Teste	21.493	27,05 (24,88)	526,41 (470,02)

3.2. Modelos de Sumarização

Os seguintes modelos de sumarização foram investigados:

Baselines. Foram utilizadas como baselines oito técnicas de ponderação de frases [Sodré and de Oliveira 2019]. As técnicas utilizadas foram: *Bushy Path*, Centralidade das Frases, Frequência de Palavras, Frequência de Entidades Nomeadas, Frequência do Termo - Frequência Inversa das Sentenças (TF-ISF), Posição das Frases, Similaridade Agregada, *TextRank*. Essas técnicas foram usadas em conjunto com uma abordagem clássica de sumarização extrativa composta por três etapas [Nenkova and McKeown 2012]:

- **Pré-processamento:** O documento ou grupo de documentos de entrada é pré-processado usando várias técnicas tradicionais de PLN, como divisão do texto em frases, palavras, lematização, identificação das classes gramaticais e reconhecimento de entidades nomeadas.
- **Ponderação das frases:** Nesta etapa, cada uma das oito técnicas de ponderação de frases é aplicada para analisar cada frase do(s) documento(s) de entrada e gerar um valor que deve refletir sua relevância para ser incluído no resumo. Todos os valores gerados são normalizados no intervalo de 0 a 1.
- **Geração de resumo:** As frases com os maiores valores de relevância geradas na etapa anterior são inseridas iterativamente no resumo até que o tamanho máximo desejado seja atingido. Uma nova frase é inserida no resumo somente

se sua similaridade de cosseno com as frases já inseridas for menor que 0,5 [Nenkova and McKeown 2012].

Sistema de PLI. Foi utilizado o sistema de PLI proposto por Gomes e Oliveira [Gomes and de Oliveira 2019] para sumarização multidocumento e uma adaptação de um sistema similar apresentado em [Oliveira et al. 2016b] para a tarefa de sumarização monodocumento.

Modelos Pré-treinados. Foram utilizados os modelos *PTT5* [Carmo et al. 2020] e *FLAN-T5* em suas arquiteturas *small*, *base* e *large*, que se diferenciam pelo tamanho da arquitetura. O *PTT5* é uma versão em português do modelo de linguagem *T5*, pré-treinado no BrWac, um grande corpus de páginas da web em português brasileiro. O *FLAN-T5* é um modelo multilíngue desenvolvido pela google que foi treinado para múltiplas tarefas de PLN [Chung et al. 2024].

LLMs. Os recentes avanços no progresso de LLMs têm impulsionado o desenvolvimento de diversas aplicações. Neste trabalho, foram utilizados os modelos: Gemma 2 9B [Team et al. 2024], o Llama 3.1 8B [Touvron et al. 2023] e os modelos Text-davinci-003, GPT-3.5 Turbo, GPT-4o e GPT-4o mini desenvolvidos pela empresa OpenAI [OpenAI 2024].

3.3. Desenho Experimental

A análise de desempenho dos modelos de sumarização foi dividida em dois experimentos. No primeiro experimento, foi utilizada somente a base de dados do RecognaSumm, sendo considerados os sistemas extrativos (*baselines* e o sistema de PLI) e foram treinados seis modelos de sumarização abstrativos baseados no *PTT5* e *FLAN-T5*. O segundo experimento foi realizado usando as bases de dados do Temário e CSTNews. Para esse experimento, foram usados os sistemas extrativos (*baselines* e o sistema de PLI), os modelos abstrativos baseados no *PTT5* e *FLAN-T5* treinados no primeiro experimento e os modelos de LLMs. Em todas as abordagens avaliadas, foi configurado o tamanho máximo de resumo para 150 palavras.

Para os métodos de *baselines* e o sistema de PLI foram usadas implementações próprias. Os modelos da OpenAI foram acessados usando a API oficial disponibilizada pela empresa. A implementação dos modelos *PTT5*, *FLAN-T5* e dos LLMs do *Gemma 2 9B* e *Llama 3.1 8B* foi baseada na biblioteca *Transformers*² e foram usados os modelos pré-treinados disponibilizados publicamente pelos autores e empresas na plataforma do Hugging Face³. Para o ajuste fino das três arquiteturas dos modelos *PTT5* e *FLAN-T5*, o tamanho máximo de entrada foi definido para 512 *tokens* e o tamanho máximo do resumo a ser gerado foi configurado para 150 *tokens*. Os modelos foram ajustados por no máximo 20 épocas, sendo utilizada a estratégia de parada antecipada com uma paciência de 5 épocas. Para evitar sobreajuste dos modelos, foi feito um monitoramento do treinamento, no qual, ao final de cada época, o modelo resultante é aplicado no conjunto de validação e é computada a medida do ROUGE-L, sendo armazenado somente o modelo com maior valor. Para a geração dos resumos, foi usado o algoritmo de decodificação do *Beam Search* com tamanho 5 de largura.

²<https://huggingface.co/docs/transformers/index>

³<https://huggingface.co/>

Baseado no trabalho de [Zhang et al. 2024], o seguinte *prompt* foi usado nos LLMs para geração dos resumos: “*Escreva um resumo em PORTUGUÊS DO BRASIL para o artigo de notícias a seguir com no MÁXIMO 150 palavras. ARTIGO: {TEXTO}.*”, onde *{TEXTO}* foi substituído pelo conteúdo completo do(s) artigo(s) de notícias.

O desempenho dos modelos foi avaliado utilizando as medidas de avaliação do *Recall-Oriented Understudy for Gisting Evaluation Longest Common Subsequence* (LCS) (ROUGE-L) [Lin 2004] e a do BERTScore [Zhang et al. 2019]. O ROUGE-L computa a maior cadeia em comum entre um resumo candidato e o resumo de referência, enquanto o BERTScore calcula a similaridade do cosseno entre dois textos usando representações de *embeddings* extraídas do modelo *Bidirecional Encoder Representations from Transformers* (BERT) [Devlin et al. 2019]. Por questões de espaço, são reportados somente a métrica do *f1-score*, que combina as métricas de precisão e revocação. Apesar de terem diversas limitações, essas medidas são alternativas válidas à realização de avaliações manuais e, conforme análise feita em Zhang et al. [Zhang et al. 2024], elas apresentaram correlação moderada com avaliações humanas na tarefa de sumarização.

4. Resultados

4.1. Experimento na base de dados do RecognaSumm

Na Tabela 2 são apresentados os resultados dos experimentos na base de dados do RecognaSumm. Analisando o desempenho dos *baselines*, pode-se observar que a técnica da *Posição das Frases* foi a que obteve os melhores resultados. Essa técnica consiste em selecionar as n primeiras frases do documento para compor o resumo até que o tamanho máximo do resumo desejado seja alcançado. Essa técnica tem sido um dos *baselines* mais competitivos para sumarização de artigos de notícias [Oliveira et al. 2016a]. O sistema baseado em PLI obteve melhor desempenho do que quase todos os *baselines*, com exceção da *Posição das Frases*. Os modelos *PTT5* e *FLAN-T5* demonstraram melhor desempenho geral do que as demais abordagens analisadas. Em especial, os melhores desempenhos neste experimento foram obtidos pelos modelos *FLAN-T5_{Large}* e *PTT5_{Large}* em ambas as medidas de avaliação. Os resultados obtidos usando a arquitetura *base* foram muito próximos às arquiteturas da *large*, sendo que eles são menores e consomem menos recursos computacionais.

Com base nos resultados, fica evidente que os modelos ajustados para sumarização abstrativos geraram resumos melhores do que as técnicas de *baselines* e que o sistema extrativo de PLI nas medidas do ROUGE-L e BERTScore. Essa superioridade demonstra a eficácia dos modelos *PTT5* e *FLAN-T5* para a tarefa de geração de resumos abstrativos. Entretanto, ao considerar o uso desses modelos, é importante levar em conta o custo computacional associado a cada um, tanto para o treinamento quanto para a geração dos resumos. Portanto, a relação custo-benefício deve ser ponderada na escolha da abordagem, especialmente em cenários com recursos computacionais limitados.

4.2. Experimento nas bases de dados do TeMário e CSTNews

A Tabela 3 apresenta os resultados do experimento nas bases de dados do TeMário e CSTNews. As abordagens avaliadas incluem os métodos de *baselines*, o sistema extrativo usando PLI, os modelos do *PTT5* e *FLAN-T5* treinados no RecognaSumm e os LLMs analisados. Os resultados obtidos neste experimento foram bastante diversificados.

Tabela 2. Resultados do experimento usando o corpus RecognaSumm.

Abordagem		ROUGE-L	BERTScore
<i>Baselines</i>	Bushy Path	0,249 (0,086)	0,691 (0,037)
	Centralidade das Frases	0,249 (0,087)	0,690 (0,038)
	Frequência de Palavras	0,240 (0,085)	0,686 (0,038)
	Freq. Entidades Nomeadas	0,242 (0,088)	0,681 (0,038)
	Posição das Frases	0,279 (0,099)	0,701 (0,040)
	Similaridade Agregada	0,249 (0,085)	0,689 (0,039)
	TextRank	0,206 (0,072)	0,674 (0,034)
	TF-ISF	0,235 (0,084)	0,684 (0,037)
<i>Extrativo</i>	Sistema PLI	0,270 (0,095)	0,694 (0,038)
<i>Abstrativos</i>	<i>PTT5_{Small}</i>	0,315 (0,125)	0,713 (0,045)
	<i>PTT5_{Base}</i>	0,337 (0,132)	0,722 (0,045)
	<i>PTT5_{Large}</i>	0,346 (0,134)	0,726 (0,046)
	<i>FLAN-T5_{Small}</i>	0,314 (0,130)	0,714 (0,045)
	<i>FLAN-T5_{Base}</i>	0,338 (0,140)	0,724 (0,048)
	<i>FLAN-T5_{Large}</i>	0,349 (0,143)	0,729 (0,048)

Na base de dados do Temário, os modelos Text-davinci-003 e GPT-4o obtiveram o melhor desempenho nas medidas do ROUGE-L e BERTScore, respectivamente. Na base do CSTNews, o *baseline* de *Posição das Frases* apresentou o melhor resultado no ROUGE-L e o GPT-3.5 Turbo no BERTScore.

Tabela 3. Resultados do experimento usando o Temário e o CSTNews.

Abordagens	Sistema	Temário		CSTNews	
		ROUGE-L	BERTScore	ROUGE-L	BERTScore
<i>Baselines</i>	Bushy Path	0,396 (0,069)	0,694 (0,024)	0,447 (0,067)	0,720 (0,029)
	Cent. das Frases	0,384 (0,063)	0,690 (0,025)	0,454 (0,066)	0,723 (0,031)
	Freq. de Palavras	0,375 (0,069)	0,686 (0,023)	0,452 (0,063)	0,721 (0,029)
	Freq. Ent. Nom.	0,389 (0,076)	0,683 (0,024)	0,434 (0,068)	0,705 (0,032)
	Posição das Frases	0,402 (0,070)	0,686 (0,022)	0,482 (0,047)	0,733 (0,024)
	Sim. Agregada	0,390 (0,070)	0,696 (0,025)	0,419 (0,050)	0,712 (0,024)
	TextRank	0,350 (0,059)	0,685 (0,021)	0,415 (0,060)	0,709 (0,027)
	TF-ISF	0,379 (0,072)	0,685 (0,024)	0,451 (0,061)	0,718 (0,032)
<i>Extrativo</i>	Sistema PLI	0,396 (0,065)	0,687 (0,023)	0,477 (0,049)	0,736 (0,033)
<i>Abstrativos</i>	<i>PTT5_{Small}</i>	0,348 (0,064)	0,679 (0,024)	0,393 (0,065)	0,713 (0,031)
	<i>PTT5_{Base}</i>	0,346 (0,062)	0,681 (0,023)	0,384 (0,055)	0,712 (0,028)
	<i>PTT5_{Large}</i>	0,339 (0,062)	0,678 (0,025)	0,385 (0,055)	0,715 (0,026)
	<i>FLAN-T5_{Small}</i>	0,241 (0,053)	0,654 (0,021)	0,304 (0,070)	0,700 (0,027)
	<i>FLAN-T5_{Base}</i>	0,242 (0,048)	0,658 (0,022)	0,290 (0,064)	0,700 (0,033)
	<i>FLAN-T5_{Large}</i>	0,225 (0,049)	0,654 (0,021)	0,294 (0,070)	0,696 (0,033)
<i>LLMs</i>	Gemma 2 9B	0,354 (0,046)	0,690 (0,020)	0,383 (0,034)	0,717 (0,023)
	Llama 3.1 8B	0,320 (0,037)	0,671 (0,019)	0,338 (0,037)	0,699 (0,026)
	Text-davinci-003	0,424 (0,075)	0,705 (0,027)	0,472 (0,048)	0,738 (0,029)
	GPT-3.5 Turbo	0,402 (0,074)	0,705 (0,025)	0,455 (0,061)	0,740 (0,025)
	GPT-4o	0,417 (0,062)	0,713 (0,025)	0,452 (0,043)	0,731 (0,023)
	GPT-4o Mini	0,402 (0,059)	0,705 (0,021)	0,444 (0,039)	0,730 (0,023)

Os métodos de *baselines*, o sistema extrativo baseado em PLI e os LLMs apresentaram resultados próximos com base nas medidas de avaliação. Por outro lado, os modelos ajustados do *PTT5* e *FLAN-T5* demonstraram desempenho inferior aos demais, especialmente os modelos do *FLAN-T5*. Esse baixo desempenho pode ser atribuído ao

fato desses modelos consistentemente gerarem resumos com tamanhos bem inferiores aos demais, mesmo sendo definido um tamanho máximo de 150 palavras. Essa característica aconteceu por conta do treinamento desses modelos no Recognasumm, que possui resumos de referência bem menores do que os do Temário e do CSTNews.

Apesar dos resultados quantitativos serem próximos, ao analisar os resumos gerados pelas abordagens extrativas e abstrativas, fica evidente que os resumos extrativos, em geral, possuem muitas informações contidas nos resumos de referências, mas os resumos possuem diversos problemas de coerência e coesão textual. Por outro lado, os resumos abstrativos são mais sucintos e, em sua maioria, apresentam uma boa qualidade textual em termos de coerência, coesão e estrutura ortográfica e gramatical. Os LLMs do Gemma 2 9B e do Llama 3.1 8B apresentaram uma tendência de terminar de forma brusca os resumos, por exemplo, no meio de uma frase. Cabe ressaltar que nenhum LLM foi ajustado para a tarefa de sumarização.

Por fim, é importante enfatizar que os LLMs, como Gemma, Llama e especialmente os modelos da OpenAI, possuem um custo consideravelmente maior do que os demais modelos avaliados neste trabalho. Essa característica deve ser considerada em aplicações práticas, na qual a relação custo-benefício é determinante. Nesse contexto, abordagens extrativas, como o sistema baseado PLI ou mesmo os *baselines*, podem oferecer uma alternativa que equilibra desempenho com menor custo computacional. Em cenários com recursos computacionais moderados, os modelos ajustados do *PTT5* e *FLAN-T5* podem ser as melhores opções.

5. Conclusões

Este trabalho apresentou uma análise comparativa de várias abordagens para sumarização automática de texto, considerando desde tradicionais métodos de ponderação de frases até modelos de linguagem de grande escala, para sumarização abstrativa e extrativa de artigos de notícias escritas em português do Brasil. Essa avaliação fez uso de três bases de dados e de duas medidas de avaliação automática comumente usadas na literatura. Os resultados obtidos demonstram que os modelos de LLMs são promissores para a tarefa de criação automática de resumos, mas são sistemas com uma alta complexidade que requerem muitos recursos computacionais. Portanto, modelos especializados para a tarefa de sumarização ou sistemas extrativos ainda podem ser opções viáveis, especialmente em cenários de poucos recursos.

Em trabalhos futuros, pretendemos expandir este trabalho visando: **(i)** analisar o desempenho de modelos de LLM de código aberto, considerando diferentes cenários de utilização, como *zero shot-learning*, *few-shot learning* e fazendo o ajuste fino desses modelos para a tarefa de sumarização; e **(ii)** realizar uma avaliação manual de um subconjunto dos resumos gerados para complementar as análises automáticas.

Agradecimentos

Os autores agradecem ao Ifes, apoio da FAPES e CAPES (processo 2021-2S6CD, nº FAPES 132/2021) por meio do PDPG (Programa de Desenvolvimento da Pós-Graduação, Parcerias Estratégicas nos Estados).

Referências

- Cardoso, P. C., Maziero, E. G., Jorge, M. L. C., Seno, E. M., Di Felippo, A., Rino, L. H. M., Nunes, M. d. G. V., and Pardo, T. A. (2011). Cstnews-a discourse-annotated corpus for single and multi-document summarization of news texts in brazilian portuguese. In *Proceedings of the 3rd RST Brazilian Meeting*, pages 88–105.
- Carmo, D., Piau, M., Campiotti, I., Nogueira, R., and Lotufo, R. (2020). Ptt5: Pre-training and validating the t5 model on brazilian portuguese data. *arXiv preprint arXiv:2008.09144*.
- Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, Y., Wang, X., Dehghani, M., Brahma, S., et al. (2024). Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.
- Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (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.
- Gomes, L. and de Oliveira, H. (2019). A multi-document summarization system for news articles in portuguese using integer linear programming. In *Anais do XVI Encontro Nacional de Inteligência Artificial e Computacional*, pages 622–633. SBC.
- Leite, D. S. and Rino, L. H. M. (2008). Combining multiple features for automatic text summarization through machine learning. In *International Conference on Computational Processing of the Portuguese Language*, pages 122–132. Springer.
- Levitin, D. J. (2014). *Organized Mind: Thinking Straight in the Age of Information Overload (9780698157224)*. Barnes & Noble.
- Lin, C.-Y. (2004). Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.
- Lin, H. and Ng, V. (2019). Abstractive summarization: A survey of the state of the art. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 9815–9822.
- Nenkova, A. and McKeown, K. (2012). A survey of text summarization techniques. In *Mining text data*, pages 43–76. Springer.
- Oliveira, H., Ferreira, R., Lima, R., Lins, R. D., Freitas, F., Riss, M., and Simske, S. J. (2016a). Assessing shallow sentence scoring techniques and combinations for single and multi-document summarization. *Expert Systems with Applications*, 65:68–86.
- Oliveira, H., Lima, R., Lins, R. D., Freitas, F., Riss, M., and Simske, S. J. (2016b). A concept-based integer linear programming approach for single-document summarization. In *2016 5th Brazilian Conference on Intelligent Systems (BRACIS)*, pages 403–408. IEEE.
- OpenAI (2024). Openai models. <https://openai.com/api/>.
- Paiola, P. H., de Rosa, G. H., and Papa, J. P. (2022). Deep learning-based abstractive summarization for brazilian portuguese texts. In Xavier-Junior, J. C. and Rios, R. A., editors, *Intelligent Systems*, pages 479–493, Cham. Springer International Publishing.

- Paiola, P. H., Garcia, G. L., Jodas, D. S., Correia, J. V. M., Sugi, L. A., and Papa, J. P. (2024). Recognasumm: A novel brazilian summarization dataset. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese*, pages 575–579.
- Pardo, T. A. S. and Rino, L. H. M. (2003). Temário: Um corpus para sumarização automática de textos. *São Carlos: Universidade de São Carlos, Relatório Técnico*.
- Sodré, L. and de Oliveira, H. (2019). Avaliando algoritmos de regressão para sumarização automática de textos em português do brasil. In *Anais do XVI Encontro Nacional de Inteligência Artificial e Computacional*, pages 634–645. SBC.
- Team, G., Mesnard, T., Hardin, C., Dadashi, R., Bhupatiraju, S., Pathak, S., Sifre, L., Rivi re, M., Kale, M. S., Love, J., et al. (2024). Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.
- Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozi re, B., Goyal, N., Hambro, E., Azhar, F., et al. (2023). Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.
- Vaswani, A. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*.
- Zhang, M., Zhou, G., Yu, W., Huang, N., and Liu, W. (2022). A comprehensive survey of abstractive text summarization based on deep learning. *Computational intelligence and neuroscience*, 2022(1):7132226.
- Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., and Artzi, Y. (2019). Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.
- Zhang, T., Ladhak, F., Durmus, E., Liang, P., McKeown, K., and Hashimoto, T. B. (2024). Benchmarking large language models for news summarization. *Transactions of the Association for Computational Linguistics*, 12:39–57.

Boosting not so Large Language Models by using Knowledge Graphs and Reinforcement Learning

William Jones Beckhauser¹, Renato Fileto¹

¹Department of Informatics and Statistics (INE)
Federal University of Santa Catarina (UFSC), Florianópolis - Brazil

{beckhauserwilliam@gmail.com, r.fileto@ufsc.br}

Abstract. *Ensuring the viability of large language models (LLMs) in situations requiring data privacy with limited on-premise resources is a significant current challenge. This work investigates how to tackle this challenge using knowledge graphs (KGs) and reinforcement learning (RL) to enhance minor LLMs by reducing non-factual responses and response gaps. We evaluated variations of GPT (4o, 4, and 3.5), Llama2 (7b, 13b, and 70b), and Llama3 (8b and 70b) for multi-label classification and information extraction, with or without KG and RL, and also fine-tuned a BERT model. Llama3 8b combined with KG and RL outperformed all other LLM models, and the fine-tuned BERT model too.*

1. Introduction

Large language models (LLMs) such as GPT [Liu et al. 2023], Llama [Gao et al. 2023], and Gemini [Team et al. 2023] are increasing their parameter count with each new release, for performance gains [Xue et al. 2024]. Nevertheless, this technology, usually available in the clouds of large private corporations, remains out of reach for many companies and projects that need to operate on local servers [Yao et al. 2024], due to high costs and regulations like the General Data Protection Law (LGPD) [Erickson 2018]. These enterprises could rely on open-source models with many parameters, but their computational requirements are too high to run on-premises [Alizadeh et al. 2023].

Nowadays, there is a subtle research movement towards smaller open-source LLMs [Shridhar et al. 2023, Shen et al. 2024], and an intense pursuit of optimization strategies. A promising direction is Retrievable Augmented Generation (RAG) using a Knowledge Graph (KG) to add relevant formal knowledge to LLMs [Pan et al. 2023]. This approach has been tested in various tasks, including fake news detection [Liu et al. 2024], text classification [Shi et al. 2023], and refined node classification in citation graphs and networks [Bruno et al. 2023, He et al. 2023]. In the biomedical domain, these solutions have been applied in recommendation systems and drug-gene interaction studies [Xu et al. 2024, Wang et al. 2023], as well as in recruiting for clinical studies [Guan et al. 2023]. However, there are still few concrete examples demonstrating consistent performance gains by using approaches like Graph-RAG [Pan et al. 2023] in typical machine learning tasks, such as multi-label classification or information extraction, especially when using open-source LLMs.

This article contributes to filling this research gap by evaluating the synergism of KGs, reinforcement learning (RL) and LLMs. We compare the performance of relatively small LLMs, like Llama2 (7b and 13b) and Llama3 (8b), with that of larger LLMs

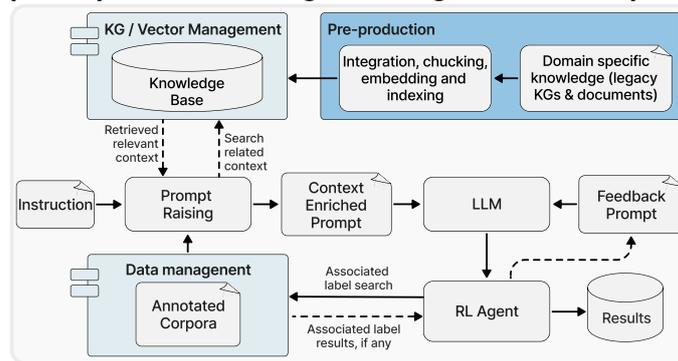
like GPT (4, 4o, and 3.5), Llama2 (70b), and Llama3 (70b), each one alone and combined with the use of KGs and RL, in two tasks: (i) multi-label classification of reviews posted by users of a food delivery app in multiple languages and their translations into English, and (ii) information extraction from invoices of different types of backhoes. We propose and evaluate alternative approaches for exploiting domain specific KGs to enrich LLM prompts with relevant context. An RL agent validates responses, restricting them to predefined labels, when available, and providing feedback to the models. It randomly validates some LLM responses with their respective labels throughout the RL process. We also fine-tuned and evaluated a BERT model for performing the same multi-label classification, on the same datasets.

The main contributions of this article are: (i) a systematic evaluation of language models, considering each LLM alone and assisted by a KG and/or an RL agent; (ii) demonstrating the superiority of smaller, open-source models, like Llama3 8b, when combined with KGs; (iii) showcasing the feasibility of feedback systems for language models; and (iv) applying LLMs combined with KGs and RL in unexplored fields.

2. Proposed Approach

Figure 1 shows the architecture of our integrated Graph-RAG and RL system for LLMs, designed to optimize responses in classification and information extraction tasks. The process starts with a instruction sent to the Prompt Raising module, supplemented by data from annotated corpora (e.g., a backhoe invoice). This module interacts with the KG/vector management component to search the Knowledge Base for relevant context by accessing the knowledge graph linked to the instruction. The retrieved context is then integrated into the prompt, which the LLM uses to generate a response. The RL Agent checks the LLM’s output against available labels(train data). If inaccuracies arise, feedback is given to the LLM, and interaction results are stored in the Results database.

Figure 1. Proposed process for using knowledge and RL to improve LLM results.



2.1. Pre-production

In the pre-production phase, we focus on constructing KGs using domain-specific structured data sources. For example, in information extraction tests related to backhoe loaders, we use tables with product descriptions segmented into products, brands, and models. These are organized into a hierarchy of classes and subclasses, with connections like “Product” connected by “offered by” to “Brand,” which in turn connects via “has”

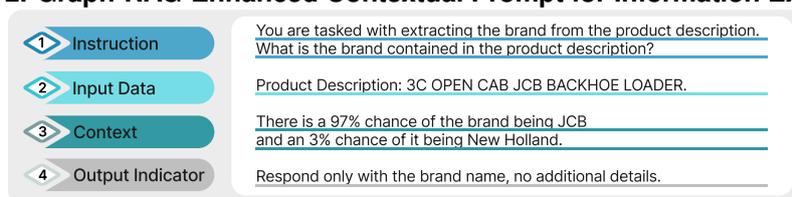
to “Model”. Once the structure is defined and validated, we generate embeddings for the classes, subclasses, and connections using the BGE model [Chen et al. 2024]. The graphs are then implemented in the Neo4j graph database, incorporating the generated embeddings. These KGs stored in Neo4j serve as our Knowledge Base.

2.2. Prompt Raising

In the "Prompt Raising" phase, the system processes three inputs: the instruction, textual descriptions on data management, and relevant information from the knowledge base. The third input is obtained via two methods: Graph-RAG Targeted, retrieving highly similar information, and Graph-RAG Comprehensive, gathering related classes and their interrelationships without filters.

Embeddings are generated from the inputs using the BGE model, the same one employed for knowledge graphs (KGs). Then, a similarity search compares these embeddings with the knowledge base using cosine similarity. The Graph-RAG Targeted method identifies records with a cosine similarity above 85%, while the Graph-RAG Comprehensive method retrieves all relevant classes, subclasses, and connections. For example, when processing "3C OPEN CAB JCB BACKHOE LOADER" in a brand-related instruction, Graph-RAG Targeted might indicate "97% probability for JCB and 3% for New Holland," while Graph-RAG Comprehensive would provide broader insights such as "The JCB brand includes models 3CX and 5CX" and "New Holland covers models B95C and B115C". Thus, as shown in Figure 2, the output of Prompt Raising consists of Instruction and Output Indicator. The Input Data represents a textual description from Data Management, and the Context, in this example, is Graph-RAG Targeted, which retrieved the data with the highest cosine similarities from our Knowledge Base.

Figure 2. Graph-RAG-Enhanced Contextual Prompt for Information Extraction.



2.3. LLM

We configure the LLM and invoke its API using an enriched prompt derived from the Prompt Raising stage. Key parameters, like temperature and output token count, are adjusted. Temperature controls prediction randomness, with lower values yielding more deterministic results and higher values increasing creativity. For classification and extraction tasks, we limit the output to fewer than 10 tokens. We employ models like Llama2 (7b, 13b, 70b) and Llama3 (8b, 70b) via the Deepinfra API, as well as GPT models (3.5, 4.0) via the OpenAI API. With the enriched prompt and optimal model settings, the API is called to perform extraction or classification. For example, for the product description “3C OPEN CAB JCB BACKHOE LOADER”, the expected response would be “JCB”.

2.4. RL Agent

The RL Agent processes the output of the LLM model by checking if there is a corresponding label in the database, as detailed in the enriched prompt. The annotated corpora

include a percentage of pre-labeled data randomly distributed, and each new LLM output is compared against these corpora. For example, if the LLM classifies a product description as "New Holland" for "JCB 3C OPEN CAB Backhoe Loader," the RL Agent searches the annotated corpora to check if there is a label. If "New Holland" is correct or if there is no existing label, the response is validated and stored; if incorrect, the agent provides feedback suggesting the correct label. This process is repeated up to five times to correct and reinforce the model’s learning. For classification tasks with predefined labels, the RL Agent adopts a two-step validation process. First, it checks if the LLM’s classification matches the predefined labels. If it doesn’t match, the agent provides feedback to align the response with the established categories. In the second step, if the classification falls within the categories, the Agent validates it against the associated label (if any).

3. Application Scenarios

3.1. Multi-label Classification of Food Delivery Reviews

In our first scenario, we analyzed a dataset of around 4,000 customer reviews from a European food delivery app, ranging from 0 to 889 characters, available in [Beckhauser and Fileto 2024]. After removing duplicates and outliers, 3,451 reviews remained. Approximately 80% are in European Portuguese, with the rest in English, Spanish, Italian, and Catalan. Given the importance of English in LLM training, we created a parallel dataset by translating all reviews into English using Googletrans, with manual corrections for about 30 reviews. We then identified key terms for each label by removing stop-words in multiple languages using nltk and spaCy and extracting frequent words with the Counter library. For sentiment analysis, we used the SiEBERT model [Hartmann et al. 2023], which showed consistent performance, even when compared to GPT-4 [Krugmann and Hartmann 2024]. Sentiment analysis results and dataset details are summarized in Table 1.

We manually built a tree-like KG to categorize reviews, distinguishing between “Product” (item-related) and “Order” (delivery/service-related). Subcategories like “Quantity issue” and “Quality issue” under “Product,” and “Delivery issue” and “Praise comment” under “Order” are further refined with specific keywords.

Table 1. Dataset review distribution by class, subclass, and sentiment.

Class	Subclass	Description	#Reviews	%	Pos.	Neg.
Product	Quality issue	Issues with food preparation, taste, or hygiene.	671	19.44	26%	74%
Product	Quantity issue	Dissatisfaction with the amount or size of the portions served.	605	17.53	28%	72%
Order	Delivery issue	Problems related to delays, wrong deliveries or missing items.	1196	34.66	26%	74%
Order	Praise comment	Positive comments about the quality of the service or product.	979	28.37	98%	2%

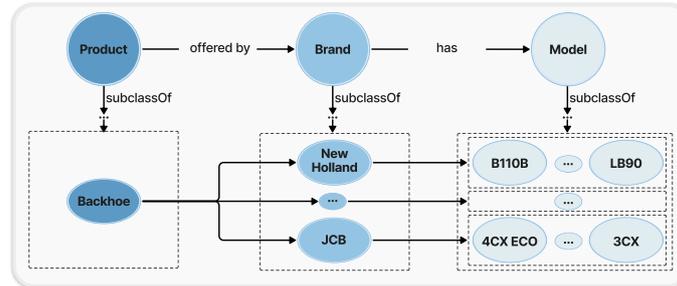
3.2. Information Extraction from Invoices

Our second application scenario involves a dataset of approximately 17,000 work machine purchase invoice descriptions, including the Mercosur Common Nomenclature (NCM)

and unit item values, provided by (blind review). The invoice descriptions range from 16 to 120 characters. Initially, we filtered the dataset using the NCM code, focusing on the first four characters, specifically "8429," which covers bulldozers, graders, excavators, and similar machinery. We then applied a keyword dictionary to identify relevant terms. Backhoe loaders appeared most frequently, with around 1,100 descriptions, becoming the primary focus of our experiments. This dataset lacked initial classifications, containing only raw invoice descriptions. To facilitate future model validation, we manually categorized the data into predefined classes such as brand, model, and specifications. A dictionary comprising brands, models, keywords, orthographic variations, acronyms, and abbreviations was used, considering possible typographical errors. Fields not covered by the dictionaries were manually completed, ensuring thorough validation of LLM outputs.

KGs for Backhoe Invoices. Figure 3 shows an extract of an ontology in KG format, centered on heavy machinery. It depicts the 'Product' concept, with 'Backhoe' as a subclass, linked to 16 brands via the "offered by" relation. Brands like 'New Holland' and 'JCB' are highlighted, each connected to specific models through the "has" relation. For instance, 'New Holland' includes models like 'B110B' and 'LB90,' while 'JCB' offers '4CX ECO' and '3CX.' In total, 68 models are represented.

Figure 3. KG extraction with concepts and relations from heavy machinery.



4. Experiments

In this section, we describe the experiments conducted for multi-label classification with customer reviews, subsection 4.1, and information extraction experiments using backhoe invoice data, subsection 4.2. All datasets and models were tested in various distinct scenarios: (1) classification or extraction using only the instruction and corpus, without providing enriched context to the LLM; (2) using only the RL Agent; (3) adding a comprehensive search in the KGs, which returns all classes, subclasses, relationships, and leafs as context; (4) using targeted context with similarity search above 85%, utilizing Graph-RAG; (5) using Graph-RAG Comprehensive with RL; (6) using Graph-RAG Targeted with RL. Additionally, for the multi-label classification experiments, we will conduct a test with embeddings and fine-tuning using BERT. A more comprehensive description of the experiments developed is available at GitHub¹.

4.1. Multilabel Classification of Customer Reviews

In this subsection, we present the experiments conducted for multi-label classification. The experiments are performed on two subsets of customer review data: the first contains

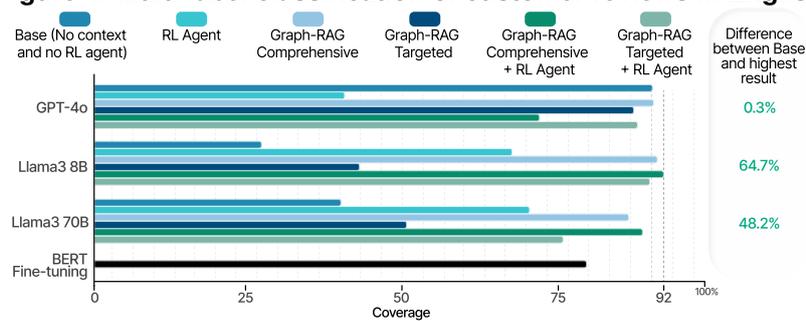
¹<https://github.com/WilliamBeckhauser/Boosting-not-so-LLM>

reviews from customers in various languages, and the second comprises the same reviews translated into English. Each dataset includes 3,451 reviews. We randomly selected 300 reviews from each label for the agent to use as a validator during the classification process, resulting in 1,200 reviews used solely for reinforcement training on the model.

The BERT experiment tokenize reviews and split them into training and testing datasets at an 80/20 ratio. We use BERT to produce embeddings and a training function with an AdamW optimizer and a linear scheduler. To optimize hyperparameters, we set up an objective function in Optuna, adjusting the learning rate, weight decay, and epochs.

English dataset: In these experiments, the Llama3 8b model, when combined with Graph-RAG Comprehensive and an RL agent, achieved a 64.7% increase in coverage compared to the “Base” experiment, the highest among all models and scenarios (Figure 4). Without Graph-RAG Comprehensive and the agent, coverage dropped drastically to 27.2%. The Llama3 8b also excelled in precision (93.3%) and F1-Score (92%) under the same conditions.

Figure 4. Multi-label classification of customer reviews in English.

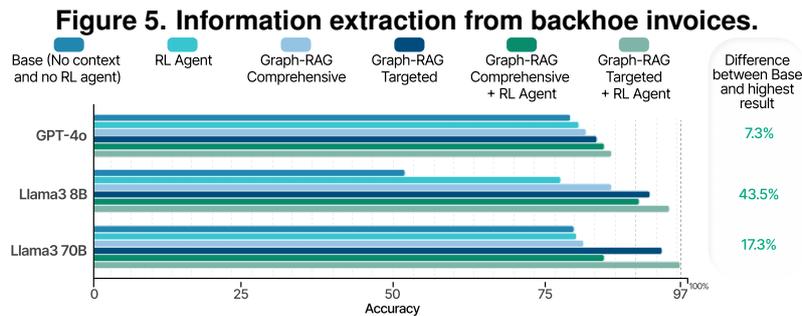


The GPT-4o model, when used solely with Graph-RAG Comprehensive, recorded 90.4% coverage, 89.1% precision, and an 89.1% F1-Score, outperforming its configuration with agents, where these metrics were 72%, 71.1%, and 71% respectively. GPT-3.5 showed moderate stability, with 81.8% coverage, 84.5% precision, and 80.6% F1-Score using Graph-RAG Comprehensive and agents. Without agents, these values only slightly declined to 81.3%, 84.1%, and 80%. The Llama2 variants underperformed, particularly the 13b version without agents. The Llama3 70b model improved in precision (92%) and F1-Score (91%) with RL agents but showed reduced performance without them. The BERT model achieved 79% precision and 76% F1-Score with general context and RL agent conditions, but still lagged behind the Llama3 models.

Multilingual dataset: The coverage improvements were more modest for GPT-4 and GPT-4o models, with only a 0.2% increase, but they retained high accuracy (around 88-89%). Notably, Llama3 70b showed strong results in both contexts, with 90.1% coverage in the multilingual setting and consistently high precision across datasets. However, in none of the scenarios did the multilingual dataset surpass the results achieved with the English dataset, highlighting a clear performance gap. Smaller models like Llama2 13b particularly struggled in both datasets, especially in multilingual tests where coverage remained low even with advanced techniques. The findings emphasize the superior adaptability of larger models like Llama3 and GPT-4 across languages, while smaller models struggle to maintain effectiveness without additional enhancements.

4.2. Information Extraction from Backhoe Invoices

Figure 5 shows that the Llama3 8b model, when operated with KGs and RL agents, displays a remarkable improvement in accuracy. Specifically, the accuracy increased from a baseline of 52.18% to 95.7% when using Graph-RAG Targeted and RL Agent, demonstrating an enhancement of 43.52%.



The Llama3 8b model achieved the highest accuracy of 95.7% and the greatest accuracy improvement among the configurations, illustrating its strong synergy with KGs and RL agents. Conversely, without these tools, its accuracy substantially decreases to the baseline of 52.18%. For the Llama3 70b model, the highest accuracy reached was 97.21% with Graph-RAG Targeted and RL Agent, showing a slight accuracy increase from its baseline of 79.93%. This model also exhibited the highest consistency across different configurations. The GPT-4o model showed improvements as well, reaching an accuracy of 86.48% with Graph-RAG Comprehensive and RL Agent, which is an increase of approximately 7.25% over its baseline of 79.23%. These results highlight the significant impact of utilizing KGs and RL agents in enhancing the performance of machine learning models, especially in tasks that involve complex document analysis such as information extraction from backhoe invoices.

4.3. Discussion

This study aligns with the growing body of research exploring the potential of LLMs to address NLP challenges. Although these models are capable of handling a wide range of tasks without the need for specialized data, in more specific cases, they show significant limitations due to the lack of fine-tuning, especially in smaller versions. LLMs face substantial limitations in their reasoning abilities, particularly when dealing with tasks involving multiple languages. In these scenarios, current LLMs still do not outperform approaches that utilize RL, whether through techniques like Proximal Policy Optimization (PPO), Trust Region Policy Optimization (TRPO), or Deep Deterministic Policy Gradient (DDPG), which require deep model adjustments, making their application considerably costly, or through RL techniques that provide textual feedback, as explored in this work. Consequently, approaches like Graph-RAG or RL with textual feedback are more viable in terms of cost and complexity.

The combination of Graph-RAG and RL, or even just one of these techniques, is more relevant for smaller models, which benefit from instructions with context and more detailed guidance, while larger models tend to perform better with more concise data or, in

some cases, no additional data at all. Even with the application of techniques like Graph-RAG, larger models maintain high effectiveness in English but exhibit performance drops when applied to multilingual datasets.

5. Related Works

Recent studies combining LLMs with KGs have focused on models like OpenAI’s GPT-3.5 and Meta’s Llama. GPT-3.5 has been applied in areas such as engineering education [Yang et al. 2023], text classification [Shi et al. 2023], and node classification in graph structures [Li et al. 2024]. GPT-4 has been used in recommendation systems and biomedical studies [Xu et al. 2024, Guan et al. 2023]. Meta’s Llama2 models have shown effectiveness in processing complex graphs, with applications in vision systems, academic databases, and digital news domains [Gouidis et al. 2024, Hu et al. 2024, Wu et al. 2024]. Chain of Thought (CoT) prompting and GNN techniques have also been integrated with LLMs for improving model interpretability and processing structured knowledge from KGs [Guan et al. 2023, Xu et al. 2023]. Techniques like PCA, UMAP, and prompt methods further integrate LLMs into the visual and structural domains of KGs, enhancing zero-shot learning [Gouidis et al. 2024, Alfasi et al. 2024]. In RL, approaches like RLHF and RLAIIF have demonstrated improvements in summarization, negotiation dialogues, and domain knowledge applications [Roit et al. 2023, Kwon et al. 2024, Mandi et al. 2023]. Although effective, RLHF and finetuning are expensive and nearly unfeasible for most experiments due to the significant computational and financial resources required [Ouyang et al. 2022, Nguyen et al. 2023]. Persistent issues like biases, toxicity, and hallucinations remain critical in both KGs and RL contexts [Gouidis et al. 2024, Xu et al. 2024, McKenna et al. 2023]. Differently from previous works, our study addresses scalability high costs associated with the use of very large model and traditional techniques fine-tuning, by combineing RAG with RL. We demonstrate the effectiveness of this approach for multi-label classification and information extraction using domain-specific KGs and datasets.

6. Conclusions and Future Work

This study demonstrates the feasibility and effectiveness of integrating LLMs with Graph-RAG to enhance multi-label classification and information extraction. Experiments conducted with variations of the GPT and Llama models, combined with the use of KGs and an RL agent, revealed significant improvements in the performance of smaller models, such as Llama3 8b, especially when combined with Graph-RAG. The combination of smaller LLMs and Graph-RAG reduces the occurrence of “hallucinations”, contributing to superior accuracy and effectiveness, even in multilingual contexts. These outcomes suggest a promising future for not so large LLM’s, especially in organizations facing data privacy constraints and computational resource limitations. As future research directions, we envision the exploitation of more diverse KGs and the investigation of RL techniques to further improve results of complex tasks. Furthermore, additional studies could apply our proposal to low resource languages, for expanding its accessibility and applicability.

Acknowledgements: This work was supported by a 2022 CNPq Universal grant, FAPESC grant 2021TR1510, the Print CAPES-UFSC Automation 4.0 Project, and indirectly by the Céos project, financed by the Public Ministry of Santa Catarina State (MPSC).

References

- [Alfasi et al. 2024] Alfasi, D., Shapira, T., and Barr, A. B. (2024). Unveiling hidden links between unseen security entities. *arXiv preprint arXiv:2403.02014*.
- [Alizadeh et al. 2023] Alizadeh, K., Mirzadeh, I., Belenko, D., Khatamifard, K., Cho, M., Del Mundo, C. C., Rastegari, M., and Farajtabar, M. (2023). Llm in a flash: Efficient large language model inference with limited memory. *arXiv preprint arXiv:2312.11514*.
- [Beckhauser and Fileto 2024] Beckhauser, W. and Fileto, R. (2024). Can a simple customer review outperform a feature set for predicting churn? In *Anais do XXXIX Simpósio Brasileiro de Bancos de Dados*, pages 117–128, Porto Alegre, RS, Brasil. SBC.
- [Bruno et al. 2023] Bruno, A., Mazzeo, P. L., Chetouani, A., Tliba, M., and Kerkouri, M. A. (2023). Insights into classifying and mitigating llms’ hallucinations. *arXiv arXiv:2311.08117*.
- [Chen et al. 2024] Chen, J., Xiao, S., Zhang, P., Luo, K., Lian, D., and Liu, Z. (2024). Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *ArXiv*, abs/2402.03216.
- [Erickson 2018] Erickson, A. (2018). Comparative analysis of the eu’s gdpr and brazil’s lgpd: Enforcement challenges with the lgpd. *Brook. J. Int’l L.*, 44:859.
- [Gao et al. 2023] Gao, P., Han, J., Zhang, R., Lin, Z., Geng, S., Zhou, A., Zhang, W., Lu, P., He, C., Yue, X., et al. (2023). Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv preprint arXiv:2304.15010*.
- [Gouidis et al. 2024] Gouidis, F., Papantoniou, K., Patkos, K. P. T., Argyros, A., and Plexousakis, D. (2024). Fusing domain-specific content from large language models into knowledge graphs for enhanced zero shot object state classification. *arXiv arXiv:2403.12151*.
- [Guan et al. 2023] Guan, Z., Wu, Z., Liu, Z., Wu, D., Ren, H., Li, Q., Li, X., and Liu, N. (2023). Cohortgpt: An enhanced gpt for participant recruitment in clinical study. *arXiv preprint arXiv:2307.11346*.
- [Hartmann et al. 2023] Hartmann, J., Heitmann, M., Siebert, C., and Schamp, C. (2023). More than a feeling: Accuracy and application of sentiment analysis. *International Journal of Research in Marketing*, 40(1):75–87.
- [He et al. 2023] He, X., Bresson, X., Laurent, T., Perold, A., LeCun, Y., and Hooi, B. (2023). Harnessing explanations: Llm-to-llm interpreter for enhanced text-attributed graph representation learning. In *ICLR*.
- [Hu et al. 2024] Hu, S., Zou, G., Yang, S., Zhang, B., and Chen, Y. (2024). Large language model meets graph neural network in knowledge distillation. *arXiv preprint arXiv:2402.05894*.
- [Krugmann and Hartmann 2024] Krugmann, J. O. and Hartmann, J. (2024). Sentiment analysis in the age of generative ai. *Customer Needs and Solutions*, 11(1):1–19.
- [Kwon et al. 2024] Kwon, D., Weiss, E., Kulshrestha, T., Chawla, K., Lucas, G. M., and Gratch, J. (2024). Are llms effective negotiators? systematic evaluation of the multi-faceted capabilities of llms in negotiation dialogues. *arXiv preprint arXiv:2402.13550*.

- [Li et al. 2024] Li, R., Li, J., Han, J., and Wang, G. (2024). Similarity-based neighbor selection for graph llms. *arXiv preprint arXiv:2402.03720*.
- [Liu et al. 2024] Liu, X., Li, P., Huang, H., Li, Z., Cui, X., Liang, J., Qin, L., Deng, W., and He, Z. (2024). Fakenewsgpt4: Advancing multimodal fake news detection through knowledge-augmented llms. *arXiv preprint arXiv:2403.01988*.
- [Liu et al. 2023] Liu, X., Zheng, Y., Du, Z., Ding, M., Qian, Y., Yang, Z., and Tang, J. (2023). Gpt understands, too. *AI Open*.
- [Mandi et al. 2023] Mandi, Z., Jain, S., and Song, S. (2023). Roco: Dialectic multi-robot collaboration with large language models. *arXiv preprint arXiv:2307.04738*.
- [McKenna et al. 2023] McKenna, N., Li, T., Cheng, L., Hosseini, M. J., Johnson, M., and Steedman, M. (2023). Sources of hallucination by large language models on inference tasks. *arXiv preprint arXiv:2305.14552*.
- [Nguyen et al. 2023] Nguyen, H. A., Stec, H., Hou, X., Di, S., and McLaren, B. M. (2023). Evaluating chatgpt’s decimal skills and feedback generation in a digital learning game. In *European Conference on Technology Enhanced Learning*, pages 278–293. Springer.
- [Ouyang et al. 2022] Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. (2022). Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.
- [Pan et al. 2023] Pan, S., Luo, L., Wang, Y., Chen, C., Wang, J., and Wu, X. (2023). Unifying large language models and knowledge graphs: A roadmap. *ArXiv*, abs/2306.08302.
- [Roit et al. 2023] Roit, P., Ferret, J., Shani, L., Aharoni, R., Cideron, G., Dadashi, R., Geist, M., Girgin, S., Hussenot, L., Keller, O., et al. (2023). Factually consistent summarization via reinforcement learning with textual entailment feedback. *arXiv preprint arXiv:2306.00186*.
- [Shen et al. 2024] Shen, W., Li, C., Chen, H., Yan, M., Quan, X., Chen, H., Zhang, J., and Huang, F. (2024). Small llms are weak tool learners: A multi-llm agent. *arXiv preprint arXiv:2401.07324*.
- [Shi et al. 2023] Shi, Y., Ma, H., Zhong, W., Tan, Q., Mai, G., Li, X., Liu, T., and Huang, J. (2023). Chatgraph: Interpretable text classification by converting chatgpt knowledge to graphs. In *2023 IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 515–520. IEEE.
- [Shridhar et al. 2023] Shridhar, K., Sinha, K., Cohen, A., Wang, T., Yu, P., Pasunuru, R., Sachan, M., Weston, J., and Celikyilmaz, A. (2023). The art of llm refinement: Ask, refine, and trust. *arXiv preprint arXiv:2311.07961*.
- [Team et al. 2023] Team, G., Anil, R., Borgeaud, S., Wu, Y., Alayrac, J.-B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A. M., Hauth, A., et al. (2023). Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.
- [Wang et al. 2023] Wang, Q., Gao, Z., and Xu, R. (2023). Graph agent: Explicit reasoning agent for graphs. *arXiv preprint arXiv:2310.16421*.

- [Wu et al. 2024] Wu, H., Zhang, Y., Han, Z., Hou, Y., Wang, L., Liu, S., Gong, Q., and Ge, Y. (2024). Quartet logic: A four-step reasoning (qlfr) framework for advancing short text classification. *arXiv preprint arXiv:2401.03158*.
- [Xu et al. 2024] Xu, D., Zhang, Z., Lin, Z., Wu, X., Zhu, Z., Xu, T., Zhao, X., Zheng, Y., and Chen, E. (2024). Multi-perspective improvement of knowledge graph completion with large language models. *arXiv preprint arXiv:2403.01972*.
- [Xu et al. 2023] Xu, H., Gao, Y., Hui, Z., Li, J., and Gao, X. (2023). Language knowledge-assisted representation learning for skeleton-based action recognition. *arXiv preprint arXiv:2305.12398*.
- [Xue et al. 2024] Xue, F., Fu, Y., Zhou, W., Zheng, Z., and You, Y. (2024). To repeat or not to repeat: Insights from scaling llm under token-crisis. *Advances in Neural Information Processing Systems*, 36.
- [Yang et al. 2023] Yang, Y., Chen, S., Zhu, Y., Zhu, H., and Chen, Z. (2023). Knowledge graph empowerment from knowledge learning to graduation requirements achievement. *Plos one*, 18(10):e0292903.
- [Yao et al. 2024] Yao, Y., Duan, J., Xu, K., Cai, Y., Sun, Z., and Zhang, Y. (2024). A survey on large language model (llm) security and privacy: The good, the bad, and the ugly. *High-Confidence Computing*, page 100211.

Genipapo – a Multigenre Dependency Parser for Brazilian Portuguese

Ariani Di Felippo^{1,2}, Norton T. Roman³, Bryan K. S. Barbosa^{1,2},
Thiago A. S. Pardo¹

¹ Núcleo Interinstitucional de Linguística Computacional (NILC)
Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo (USP)
São Carlos/SP – Brazil

²Departamento de Letras, Universidade Federal de São Carlos (UFSCar)
São Carlos/SP – Brazil

³Escola de Artes, Ciências e Humanidades, Universidade de São Paulo (USP)
São Paulo/SP – Brazil

ariani@ufscar.br, norton@usp.br, bryankhelven@ieee.org
taspardo@icmc.usp.br

Abstract. *In this article, we present a pioneer effort on building a multigenre parsing model for Brazilian Portuguese. Following the Universal Dependencies framework, we trained a current state-of-the-art model in three corpora from different text genres (journalistic, academic and user-generated content – X posts). Our experiments show that our multigenre parsing model achieves better or competitive results in relation to single-genre trained parsers.*

Resumo. *Neste artigo, apresenta-se um esforço pioneiro para o desenvolvimento de um modelo de parsing multigênero para o português brasileiro. Seguindo o projeto Universal Dependencies, treinou-se um dos modelos do estado-da-arte em três corpora gold-standard de diferentes gêneros textuais (jornalístico, acadêmico e conteúdo gerado por usuário – postagens do X). Os experimentos revelam que nosso modelo multigênero de parsing produz resultados melhores ou competitivos em relação aos modelos de gênero único.*

1. Introduction

Syntactic parsing is the task of automatically uncovering the syntactic relations among the words of a sentence, resulting in syntactic trees, which correspond to one of the first analysis levels in Natural Language Processing (NLP) [Jurafsky and Martin 2024]. This task has proved useful for several different applications, such as text simplification, information extraction, automatic summarization, and sentiment analysis, among many others.

As time goes by, parsing takes different importance degrees and attend different desires. In the beginning, it was common to have parsing as a step in NLP applications (*e.g.* grammar checking [Martins et al. 1998] and text simplification [Candido et al. 2009]). Recent advances in deep learning, distributional models, and language modeling have allowed many applications to forgo deeper linguistic analysis, but current research efforts have indicated that the inclusion of linguistic knowledge during model training or in post-processing steps (*e.g.* in neuro-symbolic approaches)

may be relevant for improving results [Zhou et al. 2020, Bai et al. 2021, Lin et al. 2021, Bölücü et al. 2023]. Moreover, given the expensive computational requirements for training the above models and the search for explicability and interpretability, linguistic analysis systems have reemerged as relevant alternatives in several research situations.

There are some well known parsers for Portuguese, including those considered classic, such as PALAVRAS [Bick 2000] and *PassPort* [Zilio et al. 2018], and more recent ones aligned to the *Universal Dependencies* (UD) project [de Marneffe et al. 2021], as *UDPipe 2* [Straka 2018] and the current state-of-the-art *Portparser* [Lopes and Pardo 2024] (with accuracy near 95% for news texts).

We propose here to move a step further in parsing for Brazilian Portuguese (BP). Using the different annotated corpora that are available in the UD initiative, and adopting a widely known parsing framework (the *Stanza* pipeline [Qi et al. 2020]), we investigate the issue of multigenre parsing, aiming at producing a parser that works well for different language writing styles, including short and usually syntactically fragmented X posts (formerly known as tweets), “daily language” of news texts and (supposedly) more refined writing of academic texts. The resulting system, named Genipapo¹ (an acronym for “multiGENre PARser for PORTuguese”), achieves better or competitive results in relation to the single-genre trained parsers, consisting in a step to unleash the potential of Portuguese text analysis tools to work on a wide variety of texts.

The rest of this article is organized as follows. Section 2 introduces the UD framework. Section 3 briefly presents the main related work in the area. The adopted resources and methodology are reported in Section 4, whereas Section 5 presents the results of our experiments. We conclude this article in Section 6.

2. The *Universal Dependencies* framework

UD [Nivre et al. 2020] is currently the most used dependency-based framework of morphological and syntactic analysis in NLP [Sanguinetti et al. 2023]. It is an attempt to standardize the annotation of morphology and syntax, proposing a “universal” annotation strategy for all languages, facilitating the development of multilingual taggers and parsers. At the time of this writing, there are already over 240 treebanks available for more than 150 languages, dealing with a variety of textual genres.

In UD, the following morphology information is considered: (i) Part-of-Speech (PoS) tags, (ii) lemmas, and (iii) features. The syntactic annotation consists of typed dependency relations (*deprels*) between words. Currently, the model has 17 PoS tags and 37 *deprels*, plus a non-fixed set of morphological features. Figure 1 shows an example of an annotated post from the *DANTEStocks* corpus [Di-Felippo et al. 2021]. The basic dependency representation is a tree, where exactly one word is the *head* of the utterance (**root**) (e.g. “assina” – “sign”), and all the remaining words depend on some other word. The labeled arcs represent the dependency relations, pointing from *heads* to their dependents. PoS tags, lemmas, and morphological features are displayed below the words in Figure 1.

¹The corresponding fruit, “Jenipapo” (with ‘J’ instead of ‘G’), is a tropical fruit, appreciated in several states of Brazil and used for different purposes, from painting to eating and preparing beverages. By adopting this inspiration for the name of our parser, we sought this symbolic connection with something rooted in the Brazilian culture and language.

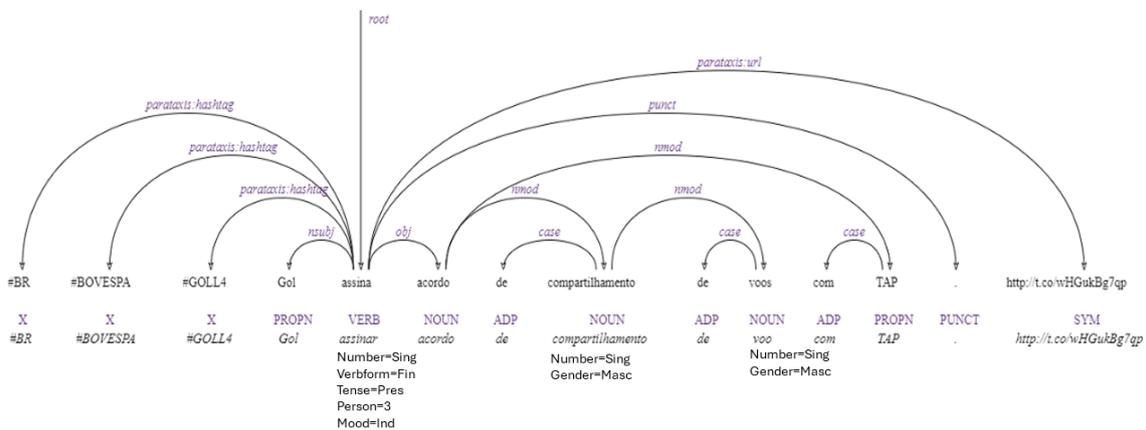


Figure 1. Example of UD morphological and syntactic annotation.

3. Related work

About the linguistic resources for training UD-parsers, there are some available datasets in BP. One of the first corpora with UD annotation for texts in standard (or canonical) Portuguese is the *UD-Portuguese-Bosque* treebank [Rademaker et al. 2017], which contains 210,958 tokens across 9,357 sentences. The Brazilian portion of this corpus consists of 4,213 well-written sentences extracted from journalistic texts. There is also *PetroGold* [Souza et al. 2021], which is a fully revised treebank that consists of academic texts from the oil and gas domain, in a total of 8,946 sentences (and 232,333 tokens). Differently from *UD-Portuguese-Bosque*, *PetroGold* is a specialised or domain-specific corpus. Besides, the UD project makes available the *UD-Portuguese-GSD* corpus [Zeman 2017]. Totalling 12,020 sentences (296,169 tokens) from news texts and blogs, it features two different textual genres, with different degree of canonicalness.

Specifically aiming at growing syntax-based resources for BP, another treebank (with genres beyond newswire texts) has been created. *Porttinari* [Pardo et al. 2021] currently includes two main genres (with others under construction): (i) news texts, representing standard written language, and (ii) user-generated content (UGC), representing informal non-canonical web language (in particular, tweets/X posts).

Concerning parsing models, some dependency UD-parsers are available for BP, specially for news texts. *UDPipe 2* [Straka 2018] is probably the most used model. Using a graph-based biaffine attention architecture, it achieves a *Labelled Attachment Score*² (LAS) of 87.04% for news texts. *Stanza* [Qi et al. 2020] is another well-known system, which uses a feature-enriched Bi-LSTM-based deep biaffine neural method. According to the results for the UD version 2.12³, *Stanza* achieves 87.75% of LAS for news texts. *UDify* [Kondratyuk and Straka 2019] is another important system. It is a semi-supervised multitask self-attention model. There is also the recently released *PortParser* [Lopes and Pardo 2024], which was built by training *UDPipe 2* with *BERTimbau* [Souza et al. 2020] on the *Porttinari-base* corpus [Duran et al. 2023a], which is part of the journalistic portion of the larger *Porttinari*⁴ [Pardo et al. 2021] treebank. The model

²This score evaluates the output of a parser by considering how many words have been assigned both the correct syntactic *head* and the correct label of the relation [Nivre and Fang 2017].

³<https://stanfordnlp.github.io/stanza/performance.html>

⁴<https://sites.google.com/icmc.usp.br/poetisa/porttinari>

achieved LAS around 95%. This LAS value brings an improvement of around 7% over some well-known existing baselines for standard written Portuguese language.

As a final example, it is important to cite the work of [Zilio et al. 2018]. However delivering lower results than those by more recent works, the authors compared some previous and classical parsing methods for BP. The authors reported that the best model (called *PassPort*) achieved LAS of 85.21% in the UD corpus. In an additional small scale evaluation, the *PassPort* was manually compared to PALAVRAS, using a single corpus of 90 sentences (1,295 tokens), randomly selected from three different genres, to wit, literature, news texts and subtitles. The systems achieved similar results for dependency parsing, with a LAS of 85.02% for *PassPort* against 84.36% for PALAVRAS.

4. Materials and methods

Given the objective of building a multigenre UD parser for BP, three corpora, belonging to three different genres, build our materials.

Our first corpus, *DANTEStocks* [Di-Felippo et al. 2021], comprises 4,048 tweets (with 81,048 tokens) from the stock market domain automatically collected during 2014 (which limits each post to 140 characters). The corpus was built by fetching messages containing a ticker⁵ of one of the 73 stocks that composed Ibovespa at that time [da Silva et al. 2020]. *DANTEStocks* presents a combination of standard and non-standard written language, as well as speech marks, domain specific vocabulary and medium (Twitter) features. The dependency relations of the corpus were annotated in two semiautomatic stages [Barbosa 2024]. First, a reference subcorpus of 1,000 tweets was annotated using *UDPipe 2*, which had been trained on *UD-Portuguese-Bosque* and was chosen because it is easily available for use online and offers reliable performance. This subcorpus was then manually revised before being designated as a gold standard. The rest of the corpus was then annotated by customizing Stanza for *DANTEStocks*. We used the combined *Portinari-base* and reference subcorpus as the initial training set for *Stanza*. The resulting parsing model was used to automatically annotate a new (first) package of data (out of the remaining 3,048 tweets). The first package was manually revised and incorporated to the previous dataset, being used to start a new training run of *Stanza*. This cycle of training iteration continued incrementally until the last (in a total of 6) package was annotated/revised. Regarding LAS, the final score (6th run) achieved 94.62%, increasing 0.76% from the first run score of 93.86%.

The second corpus, *PetroGold* [de Souza and Freitas 2023], is a gold-standard treebank for the oil and gas (O&G) domain. It integrates the *Petrolês* corpus, which is a collection of academic and technical documents from public agencies such as Petrobras and “Agência Nacional do Petróleo, Gás Natural e Biocombustíveis” (ANP) [Gomes et al. 2021]. *PetroGold* is composed of 19 academic texts (theses and dissertations), with a total of 9,127 sentences and 253,640 tokens. The syntactic annotation of *PetroGold* also followed a semiautomatic approach. Specifically, four experts were responsible for reviewing the output of a customized version of *Stanza*, trained on the combination of *UD-Portuguese-Bosque* (v.2.6) and a small collection of data from the O&G domain. Through an intrinsic evaluation using a model created by the *UDPipe*

⁵A five or six-character alphanumeric string that represents a type of stock from a company, such as “PETR4” for Petrobras’ preferred stock.

tool, the corpus achieves 88.53% of LAS. For NLP purposes, the corpus is subdivided into three subsets. The subsets have 7,170, 737 and 1,039 sentences for training (80%), validation (8%) and test (12%).

Our final corpus, *Porttinari-base* [Duran et al. 2023a], is the gold-standard (*i.e.* fully manually annotated and revised) journalistic subcorpus of *Porttinari*, which is composed of 8,418 sentences (168,080 tokens) selected from *Folha de São Paulo* newspaper. The *Porttinari-base* annotation process started with an automatic annotation by *UDPipe 2* using the *UD-Portuguese-Bosque* corpus, which achieved 87% accuracy (in terms of LAS). Next, the dependency relations were manually revised in detail following an annotation manual containing specific guidelines for BP [Duran 2022]. *Porttinari-base* is also subdivided into training, validation and test subsets. The subsets have 5,893, 842 and 1,683 sentences in the train (70%), dev (10%) and test (20%) files, respectively.

For developing our parser, we employed the *Stanza* pipeline, which was trained and evaluated across different corpora. Since both *PetroGold* and *Porttinari-base* corpora already come subdivided in train, validation and test sets, we first set apart their test sets to ensure they would only be used for final evaluation purposes. After this, we unified each corpus’ train and validation sets to build a larger training set for each, which was then used in our experiments. Next, we randomly split (from a uniform distribution), *DANTEStocks* in training and test sets, following the same principle of keeping the test set strictly for final testing. Table 1 details each set size across the corpora.

Table 1. Size and proportion of train and test sets across corpora.

<i>Corpus</i>	<i>Train</i>		<i>Test</i>		<i>Unit</i>
	<i>Units</i>	<i>Proportion</i>	<i>Units</i>	<i>Proportion</i>	
DANTEStocks	3,643	90%	405	10%	tweet
UP-Portuguese-PetroGold	7,907	88%	1,039	12%	sentence
Porttinari-base	6,735	80%	1,683	20%	sentence

To assess the model’s performance across different genres, we combined the training sets from all three corpora to create a fourth, unified training set, along with a corresponding test set. A grid search was conducted for hyperparameter optimization, focusing on batch size (2000, 3000, 4000, and 5000) and dropout rate (0.2, 0.3, and 0.4), since *Stanza* does not natively support learning rate adjustments. Next, we ran 5-fold cross-validation with grid search (using the above mentioned grid) at each of the four training sets⁶, whereby each set was further split in five subsets, with four being used to train the model, and the fifth one being held for validation purposes. This subdivision procedure is repeated five times. We then selected, for each training set, the hyperparameters that produced the highest LAS value across the validation sets during cross-validation.

Having the best set of hyperparameters for each of the four corpora (*DANTEStocks*, *PetroGold*, *Porttinari-base* and their union), we retrained the model at each corpus training set, varying its random seed (42, 123, 456, 789 and 101,112), thereby changing the model’s initial configurations. To do so, *PetroGold*’s and *Porttinari-base*’s training sets were split back into their original training and validation sets, whereas *DANTEStocks*’ training set was randomly split into training and validation sets, so that the entire *DANTEStocks* corpus would contain 10% of the data for test, 10% for validation and 80% for

⁶*I.e.* each corpus’ individual training set and the largest set built from the union of these training sets.

training purposes. The best performance model, across all seeds, was then selected for each corpus. As a final step, all four models were tested and compared using the previously separated test sets, which had been reserved exclusively for this final evaluation.

5. Results and discussion

Tables 2 and 3 present the results of our model, when trained in each corpus’ training set (rows in the tables), and tested at the different test sets of the experiment. Table 2 refers to model results in terms of LAS, whereas Table 3 presents the results in terms of *Unlabelled Attachment Score*⁷ (UAS). In the tables, the “Genipapo” lines refer to the model trained in all of the available training sets, *i.e.* our multigenre model, while the “All together” columns refer to the union of all test sets.

Table 2. Model’s LAS (%) at each corpus’ test set.

Training set	Test set			
	Porttinari-base	DANTEStocks	PetroGold	All together
Porttinari-base	94.82	66.10	87.47	88.48
DANTEStocks	87.61	91.95	83.68	86.48
PetroGold	86.74	61.30	95.33	87.30
DANTEStocks + Portt.-base	94.91	92.67	87.94	91.90
DANTEStocks + PetroGold	87.66	91.85	84.10	86.66
Porttinari-base + PetroGold	94.92	66.75	95.29	91.84
Genipapo	94.94	92.69	95.11	94.75

Table 3. Model’s UAS (%) at each corpus’ test set.

Training set	Test set			
	Porttinari-base	DANTEStocks	PetroGold	All together
Porttinari-base	95.88	75.55	90.38	91.27
DANTEStocks	90.36	93.98	87.51	89.63
PetroGold	89.69	71.45	95.84	90.15
DANTEStocks + Portt.-base	95.95	94.39	90.97	93.86
DANTEStocks + PetroGold	90.26	93.97	88.04	89.76
Porttinari-base + PetroGold	95.91	76.03	96.00	93.67
Genipapo	95.97	94.42	95.81	95.73

We see that each model trained in isolation produces the best results for its corresponding genre. For instance, considering LAS, training with *Porttinari-base* produced the best results for the test set of *Porttinari-base* (94.82%) and worse results for *DANTEStocks* (66.10%) and *PetroGold* (87.47%). This pattern holds across the genres, where the isolated models consistently perform best when tested on the same genre they were trained on. More interestingly, *Genipapo*, our multigenre parser, outperforms the single-genre trained parsers for 2 of the genres (news texts and X posts), but not for the academic genre. The differences, however, are minimal (less than 1%), suggesting that they could be due to random fluctuation rather than statistically significant differences.

When combining all test sets (“All together” columns in the tables), *Genipapo* delivers the best results, achieving a 7% improvement in LAS and nearly 5% in UAS

⁷UAS indicates the accuracy of the *head* ignoring the relation’s name (deprel) [Nivre and Fang 2017].

compared to the second-best results from single-genre parsers, and a 3% LAS and 2% UAS improvement over parsers trained on pairs. This suggests that *Genipapo* may be the more suitable choice for processing texts from varied sources, such as diverse web content.

By looking at the results produced by *Genipapo*, when tested on each corpus separately, we see some common mistakes between pairs of *deprels*. One of the most common errors across the three corpora was the confusion between **obl** and **nmod**. This result does not come as a total surprise, since the classification of a nominal as an adverbial adjunct (**obl**) or as a nominal modifier (**nmod**) was already reported in the literature as a challenge for parsing standard Portuguese (and also for humans in some situations), such as in journalistic and academic texts [Duran et al. 2023b, Souza et al. 2021]. Once this phenomenon is also observed in *DANTEStocks*' UGC, this difficulty seems to be unrelated to the degree of "canonicalness" of the corpus. The pairs **acl** (adnominal clause) and **advcl** (adverbial clause) and **obj** (the second argument of a verb) and **nsubj** (a nominal subject) show a relevant confusion only in the standard language corpora. The confusion between **acl** and **advcl** seems to be a case of ambiguity that requires semantic knowledge to be solved, and the confusion between **obj** and **nsubj** occurs when the candidate to the subject is at the right of the verb, since noun phrases at the right can be either object or subject in Portuguese [Duran et al. 2023b].

When comparing the errors of *Genipapo* at each *deprel*, we see the model making a higher number of wrong **root** predictions in *DANTEStocks*, given its error rate of 7.7% against 2.0% in *Porttinari-base* and 0.9% in *PetroGold*. This might be due to the linguistic phenomena of tweets that bring some difficulty to the syntactic annotation of the **root**. Another interesting observation relates to **parataxis**, which is one of the most frequent tag in our UGC corpus (708 cases), but not in the remaining corpora. The relatively low error rate in *DANTEStocks* (9.3%) indicates that this *deprel* has been well learned by *Genipapo* in UGC. Moreover, we could see that the *deprel* tags most wrongfully predicted due to under representation in *Porttinari-base* and *DANTEStocks* are the same: **reparandum**, **dislocated** and **orphan**. The first two tags do not occur in *PetroGold*, and the only two occurrences of **orphan** in this corpus were wrongly predicted.

As a way to compare *Genipapo*'s performance with that by a state-of-the-art model, we also run *Portparser* in the same testing sets as *Genipapo* (Table 4). We note that the training, validation, and test splits of the *Porttinari-base* used by *Portparser* differ from those publicly available and used in our experiments with *Genipapo*. This discrepancy means that some sentences present in the test sets of *Porttinari-base* and the unified set (All together) may have been included in the training or validation sets of *Portparser*, artificially boosting its LAS and UAS scores. Despite this, *Genipapo* outperformed *Portparser* across all testing sets except for the *Porttinari-base* test set. In terms of LAS, *Genipapo* showed significant improvements over *Portparser* on the *DANTEStocks* test set (92.69% vs. 64.45%), the *PetroGold* test set (95.11% vs. 86.74%), and the combined test set (94.75% vs. 89.51%). However, *Portparser* performed better on the *Porttinari-base* test set (98.06% vs. 94.94%). The same pattern is observed in UAS scores, where *Genipapo* outperformed *Portparser* on *DANTEStocks* (94.42% vs. 75.81%), *PetroGold* (95.81% vs. 90.50%), and the combined test set (95.73% vs. 92.62%). Nevertheless, *Portparser* achieved higher UAS on *Porttinari-base* (98.58% vs. 95.97%).

Table 4. *Portparser’s* LAS and UAS at each corpus’ test set.

Test set	LAS (%)	UAS (%)
Porttinari-base	98.06	98.58
DANTEStocks	64.45	75.81
PetroGold	86.74	90.50
All together	89.51	92.62

6. Final remarks

In this paper, we introduced *Genipapo*, a multigenre UD-parser for Brazilian Portuguese, and showed that it had better or competitive performance in relation to genre specific trained parsers. Future work includes (i) to extend *Genipapo’s* training to other genres and domains, such as audio transcriptions, literary texts, and tweets related to the COVID-19 pandemic, whose corpora are still under construction, and (ii) to explore different parsing strategies and pipelines.

More details about this work may be found at the POeTiSA project web portal: <https://sites.google.com/icmc.usp.br/poetisa/>.

Acknowledgements

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 2019/07665-4) and by the IBM Corporation. The project was also supported by the Ministry of Science, Technology and Innovation, with resources of Law N.8.248, of October 23, 1991, with in the scope of PPI-SOFTEX, coordinated by Softex and published as Residence in TIC13, DOU01245.010222/2022-44.

References

- Bai, J., Wang, Y., Chen, Y., Yang, Y., Bai, J., Yu, J., and Tong, Y. (2021). Syntax-BERT: Improving pre-trained transformers with syntax trees. In *Proceedings of the 16th Conference of the EACL*, p. 3011–3020.
- Barbosa, B. K. d. S. (2024). Descrição sintático-semântica de nomes predicadores em tweets do mercado financeiro em português. Master’s thesis, Programa de Pós-Graduação em Linguística, Universidade Federal de São Carlos.
- Bick, E. (2000). *The Parsing System “Palavras”*. *Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. University of Aarhus, Århus.
- Bölücü, N., Rybinski, M., and Wan, S. (2023). Investigating the impact of syntax-enriched transformers on quantity extraction in scientific texts. In *Proceedings of the 2nd Workshop on Information Extraction from Scientific Publications*, p. 1–13, Bali.
- Candido, A., Maziero, E., Specia, L., Gasperin, C., Pardo, T., and Aluisio, S. (2009). Supporting the adaptation of texts for poor literacy readers: a text simplification editor for Brazilian Portuguese. In *Proceedings of the 4th Workshop on Innovative Use of NLP for Building Educational Applications*, p. 34–42, Boulder, Colorado.
- da Silva, F. J. V., Roman, N. T., and Carvalho, A. M. B. R. (2020). Stock market tweets annotated with emotions. *Corpora*, 15(3):343–354.

- de Marneffe, M.-C., Manning, C. D., Nivre, J., and Zeman, D. (2021). Universal Dependencies. *Computational Linguistics*, 47(2):255–308.
- de Souza, E. and Freitas, C. (2023). Explorando variações no tagset e na anotação universal dependencies (ud) para português: Possibilidades e resultados com base no treebank petrogold. In *Anais do XIV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, p. 125–134, Porto Alegre, RS, Brasil. SBC.
- Di-Felippo, A., Postali, C., Ceregatto, G., Gazana, L., Silva, E., Roman, N., and Pardo, T. (2021). Descrição preliminar do corpus DANTEStocks: diretrizes de segmentação para anotação segundo Universal Dependencies. In *Anais do XIII Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, p. 335–343, Porto Alegre, RS, Brasil. SBC.
- Duran, M., Lopes, L., Nunes, M. d. G. V., and Pardo, T. A. S. (2023a). The dawn of the Portinari multigenre treebank: introducing its journalistic portion. In *Proceedings of the XIV Brazilian Symposium in Information and Human Language Technology (STIL)*, p. 115–124, Porto Alegre, RS, Brasil. SBC.
- Duran, M., Nunes, M. d. G. V., and Pardo, T. A. S. (2023b). Construções sintáticas do português que desafiam a tarefa de parsing: uma análise qualitativa. In *Proceedings of the 2nd Universal Dependencies Brazilian Festival (UDFest-BR)*, p. 424–433, Porto Alegre, RS, Brasil. SBC.
- Duran, M. S. (2022). Manual de anotação de relações de dependência - versão revisada e estendida: orientações para anotação de relações de dependência sintática em língua portuguesa, seguindo as diretrizes da abordagem Universal Dependencies (UD).
- Gomes, D. S. M., Cordeiro, F. C., Consoli, B. S., Santos, N. L., Moreira, V. P., Vieira, R., Moraes, S., and Evsukoff, A. G. (2021). Portuguese word embeddings for the oil and gas industry: Development and evaluation. *Computers in Industry*, 124:103347.
- Jurafsky, D. and Martin, J. H. (2024). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models*. 3rd edition. Online manuscript released August 20, 2024.
- Kondratyuk, D. and Straka, M. (2019). 75 languages, 1 model: Parsing Universal Dependencies universally. 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)*, p. 2779–2795, Hong Kong, China. ACL.
- Lin, Y., Wang, C., Song, H., and Li, Y. (2021). Multi-head self-attention transformation networks for aspect-based sentiment analysis. *IEEE Access*, 9:8762–8770.
- Lopes, L. and Pardo, T. (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*, p. 401–410, Santiago de Compostela, Galicia/Spain. ACL.
- Martins, R. T., Hasegawa, R., Nunes, M. d. G. V., Montilha, G., and Oliveira, O. N. (1998). Linguistic issues in the development of regra: A grammar checker for brazilian portuguese. *Natural Language Engineering*, 4(4):287–307.

- Nivre, J., de Marneffe, M.-C., Ginter, F., Hajič, J., Manning, C. D., Pyysalo, S., Schuster, S., Tyers, F., and Zeman, D. (2020). Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, p. 4034–4043, Marseille, France. ELRA.
- Nivre, J. and Fang, C.-T. (2017). Universal Dependency evaluation. In *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017)*, p. 86–95, Gothenburg, Sweden. ACL.
- Pardo, T., Duran, M., Lopes, L., Felippo, A., Roman, N., and Nunes, M. (2021). Porttinari - a large multi-genre treebank for brazilian portuguese. In *Proceedings of the XIII Brazilian Symposium in Information and Human Language Technology*, p. 1–10, Porto Alegre, RS, Brasil. SBC.
- Qi, P., Zhang, Y., Zhang, Y., Bolton, J., and Manning, C. D. (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*, p. 101–108, Online. ACL.
- Rademaker, A., Chalub, F., Real, L., Freitas, C., Bick, E., and de Paiva, V. (2017). Universal Dependencies for Portuguese. In *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017)*, p. 197–206, Pisa, Italy. Linköping University Electronic Press.
- Sanguinetti, M., Bosco, C., Cassidy, L., and et al. (2023). Treebanking user-generated content: a ud based overview of guidelines, corpora and unified recommendations. *Language Resources Evaluation*, 57:493–544.
- Souza, E., Silveira, A., Cavalcanti, T., Castro, M., and Freitas, C. (2021). Petrogold – corpus padrão ouro para o domínio do petróleo. In *Anais do XIII Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, p. 29–38, Porto Alegre, RS, Brasil. SBC.
- Souza, F., Nogueira, R., and Lotufo, R. (2020). Bertimbau: Pretrained bert models for brazilian portuguese. In *Intelligent Systems*, p. 403–417, Cham. Springer International Publishing.
- Straka, M. (2018). UDPipe 2.0 prototype at CoNLL 2018 UD shared task. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, p. 197–207, Brussels, Belgium. ACL.
- Zeman, D. e. a. (2017). CoNLL 2017 shared task: Multilingual parsing from raw text to Universal Dependencies. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, p. 1–19, Vancouver, Canada. ACL.
- Zhou, J., Zhang, Z., Zhao, H., and Zhang, S. (2020). LIMIT-BERT: Linguistics informed multi-task BERT. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, p. 4450–4461.
- Zilio, L., Wilkens, R., and Fairon, C. (2018). Passport: A dependency parsing model for portuguese. In *Computational Processing of the Portuguese Language*, p. 479–489, Cham. Springer International Publishing.

Toxic Text Classification in Portuguese: Is LLaMA 3.1 8B All You Need?

Amanda S. Oliveira¹, Pedro H. L. Silva², Valéria de C. Santos²,
Gladston Moreira², Vander L. S. Freitas², Eduardo J. S. Luz²

¹BLIP

30.130-174 – Belo Horizonte – MG – Brazil

²Computing Department – Federal University of Ouro Preto (UFOP)

35.400-000 – Ouro Preto – MG – Brazil

amanda.oliveira@blip.ai

{silvap, valeriacs, gladston, vander.freitas, eduluz}@ufop.edu.br

Abstract. *The recognition of toxic and hate speech on social media platforms is important due to the significant risks posed to users and the digital ecosystem. Current state-of-the-art models, such as BERTimbau, have set benchmarks for Portuguese text classification, yet challenges remain in accurately detecting toxic content. This paper investigates the effectiveness of fine-tuning a smaller, open-source decoder-only model, LLaMA 3.1 8B 4bit, for this task. We propose an iterative prompt evolution method to optimize the model’s performance. Our results demonstrate that fine-tuning significantly enhances the LLaMA model’s F1-score from 0.61 to 0.75, surpassing BERTimbau in precision and matching the performance of the GPT-4o mini. However, the approach depends on the quality of the language models used for prompt evolution, highlighting the need for further research to enhance robustness in this area.*

1. Introduction

The task of recognizing toxic and hate speech has gained substantial attention in recent years, particularly with the surge of user-generated content on social media platforms. As these platforms increasingly shape public discourse, the proliferation of harmful content presents significant risks to both individual users and the broader digital environment. Consequently, the need for effective moderation tools has escalated, driving research toward automated solutions capable of operating at scale.

Current state-of-the-art methods for automated toxic content classification predominantly leverage transformer-based architectures, with encoder-only models being the most common. Within the Portuguese language context, BERTimbau has emerged as a leading approach [Souza et al. 2020], demonstrating superior performance in various NLP tasks, including emotion classification [Hammes and de Freitas 2021], toxic speech detection [da Rocha Junqueira et al. 2023, Oliveira et al. 2023], news clustering [Pereira and da Silva 2023], among other tasks [dos Santos and Paraboni 2023, Serras and Finger 2021]. The BERTimbau ability to capture subtle nuances in Portuguese expressions has set a high standard in the field, making it the benchmark for multi-class classification tasks. However, despite its effectiveness, the problem of accurately classifying toxic content remains an open challenge, particularly in the diverse and evolving landscape of online discourse.

Recent advancements have shifted towards decoder-only models, such as LLM2Vec [BehnamGhader et al. 2024] and NV-Embed [Lee et al. 2024], which have shown promising results across multiple languages. Notably, OpenAI ChatGPT [OpenAI et al. 2024]¹, a large decoder-only language model, has demonstrated competitive performance in this domain [Oliveira et al. 2023]. The emergence of open-source models, like the Meta LLaMA family of models [Dubey et al. 2024], further compels a reexamination of existing methodologies, raising research questions about the potential of these newer models.

Building on these recent developments, this work explores the capabilities of decoder-only models, specifically focusing on the LLaMA 3.1 8B 4bit model [Dubey et al. 2024]. This model is particularly compelling due to its open-source nature, benchmark performance, and relatively smaller size, making it well-suited for fine-tuning specialized tasks such as toxic content classification in Portuguese. The key research questions guiding this investigation are RQ1: Can a fine-tuned LLaMA 3.1 8B 4bit model achieve or surpass the performance of GPT-4o mini in classifying toxic content in Portuguese? RQ2: Can this model outperform the current state-of-the-art BERTimbau-based approach in the same task? To address these questions, we propose a heuristic approach that utilizes a larger LLM (GPT-4o-mini) to refine the prompts employed by a smaller LLM, thereby automating prompt engineering. The optimal prompt is then used to fine-tune the LLaMA 3.1 8B 4bit model for toxic content classification in social media, using the TolDBr dataset - a large public dataset on this task [Leite et al. 2020]. Our results show that the fine-tuned LLaMA 3.1 8B 4-bit model, operating in zero-shot classification mode, outperforms the BERTimbau-based model regarding precision and is on par with GPT-4o mini.

2. Materials and Methods

Although the primary focus of this work is to investigate the performance of a small and open-source language model (with only 8B parameters) for the task of toxic text detection in Portuguese, the choice of prompt is a significant challenge. The quality of the prompt heavily influences the LLM’s performance [Brown et al. 2020]. Therefore, this work proposes a straightforward approach to evolving prompts, ultimately using the best prompt identified for fine-tuning the model.

The following subsections describe the dataset selected for benchmarking, which is a large and popular dataset by Portuguese language standards for this task. Additionally, an outline of the methodology for selecting the best prompt and the approach used for fine-tuning the model.

2.1. Told-Br dataset

We employed the ToLD-br dataset, developed in [Leite et al. 2020] for training and testing the models used in this study. This dataset contains 21,000 tweets, annotated in a binary manner as “toxic” or “non-toxic”. Additionally, the tweets are also classified into different categories of toxicity, such as LGBTphobia, insults, racism, misogyny, and xenophobia.

In this study, we focused on the binary classification between “toxic” and “non-toxic”, using the corresponding annotations to train and test our models. The dataset was

¹<https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/>

divided in a stratified manner, with 80% of the tweets allocated to the training set and the remaining 20% to the test set.

2.2. Prompt Engineering: Iterative Prompt Refinement

The challenge in using large language models (LLMs) for zero-shot classification lies in identifying the most effective prompt. This study proposes a heuristic to iteratively refine prompts using a larger LLM, intending to enhance classification accuracy in a smaller LLM.

Our approach draws on previous research, mainly works by [Oliveira et al. 2024] and [Oliveira et al. 2023], which advocate for using in-context learning for social media post classification. While these studies explore both zero-shot and few-shot modalities, our focus remains exclusively on the zero-shot scenario.

Given that LLMs have been shown to function effectively as black-box optimizers [Zheng et al. 2023] and are viable alternatives to mutation and crossover operations in genetic algorithms [Lehman et al. 2023, Meyerson et al. 2023], we draw inspiration from the work presented in [Guo et al. 2024] to propose a simplified algorithm for evolving prompts tailored explicitly to the task of toxic speech detection in Portuguese.

The methodology is structured as Figure 1 illustrates: Initially, a population of prompts is initialized, each one specifically designed to classify social media posts as toxic or non-toxic. The prompts then undergo a selection process, retaining only the top-performing ones based on evaluation metrics. Next, operations to evolve the prompt are applied utilizing an instruction to a larger LLM, such as GPT-4, which assists in generating new variations by recombining elements of existing successful prompts. This process is iteratively refined to enhance the quality of the prompts. Finally, the optimal prompt from this cycle is used to fine-tune the model. Algorithm 1 provides a pseudo-code overview of these steps.

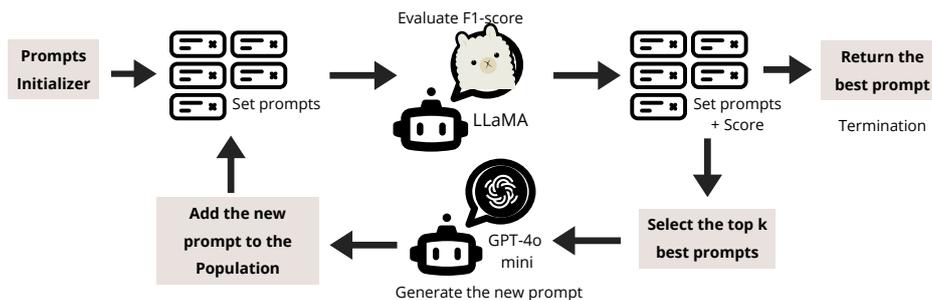


Figure 1. The heuristic iterative prompt evolution process begins with an initial set of prompts, which are evaluated using the LLaMA model based on their F1-scores. The top-performing prompts are then selected, and GPT-4o mini generates new prompts. These new prompts are added back to the population, and the process repeats. The best prompt from this iterative cycle is ultimately selected for further use. All prompts and instructions used in this study were written in Portuguese.

2.3. LLM Fine-Tuning Methodology

This methodology fine-tunes the model using a quantized version to enhance memory efficiency and speed. Parameter-Efficient Fine-Tuning (PEFT) [Houlsby et al. 2019,

Algorithm 1 Simplified Prompt Evolution

```

1: Function InitializePopulation(InitialPrompts)
2:   Population  $\leftarrow$  []
3:   for each prompt in InitialPrompts do
4:     Evaluate prompt with Llama 3.1 8B, using F1-score
5:     Add prompt and its score to Population
6:   end for
7:   return Population
8: Function GenerateNewPrompt(PromptsAndScores)
9:   PromptsText  $\leftarrow$  Concatenate each prompt and its F1-Score from PromptsAndScores with line breaks
10:  SystemInstruction  $\leftarrow$  “You are an assistant that helps improve AI prompts. You should always
    generate a new prompt, using different words or varying lengths, never repeating the same prompt.
    Generate ONLY the prompt, without comments or explanations”.
11:  Instruction  $\leftarrow$  “You are evolving a prompt for another LLM. Based on the following prompts and
    their respective F1-scores, generate a new prompt optimized for the task of classifying hate speech”.
12:  ChatGPTInput  $\leftarrow$  SystemInstruction + Instruction + PromptsText
13:  NewPrompt  $\leftarrow$  Call ChatGPT API with ChatGPTInput
14:  return NewPrompt
15: Function Main()
16:  InitialPrompts  $\leftarrow$  Define initial set of prompts
17:  Population  $\leftarrow$  InitializePopulation(InitialPrompts)
18:  for each epoch in range(NumEpochs) do
19:    TopKPrompts  $\leftarrow$  Select top ‘k’ prompts from Population, based on F1-scores
20:    PromptsAndScores  $\leftarrow$  Collect scores and prompts from TopKPrompts
21:    NewPrompt  $\leftarrow$  GenerateNewPrompt(PromptsAndScores)
22:    Evaluate NewPrompt with Llama 3.1 8B using F1-score
23:    Add NewPrompt and its score to Population
24:  end for
25:  BestPrompt  $\leftarrow$  Select best performing prompt from Population
26:  return BestPrompt

```

Hu et al. 2021] and QLoRA [Detmeters et al. 2023] techniques reduce model complexity, focusing on optimizing QKV projections and Feed Forward Layers. Training data is divided into training and validation sets. Specific prompts, structured as Alpaca prompts, align the model with toxic content classification objectives in Portuguese.

Alpaca Prompt Example:

Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request.

Instruction: You are analyzing a social media post. If the text contains hate speech, offenses, aggressions, insults, swear words, or any form of toxicity, respond only with ‘yes’. If it is appropriate and non-toxic, respond ‘no’.

Input: Tweet: every time the weather changes, my throat gets inflamed...dammit!

Response: no.

3. Experimental Setup and Results

Two key experiments are conducted to address the research questions posed in this study. First, Experiment #1 is designed to identify the best prompt. Using this prompt, the LLaMA 3.1 8B model is fine-tuned with the training data from the ToLD-Br dataset (Experiment #2). Four versions of the model are fine-tuned, varying parameters related to PEFT/QLoRA. An additional experiment is proposed to evaluate the performance of using a fine-tuned GPT-4o mini model within the same scenario. All experiments are performed in a Google Colab environment, utilizing an A100 GPU. Source code is available at <https://github.com/oliveiraamanda/ToxicSpeech-Llama-STIL-2024>.

3.1. Experiment #1 - Prompt Engineering

To evolve the prompts using the iterative algorithm proposed here, it is essential first to define the cost function to be minimized. The F1-score of binary classification on a partition of the training data is selected as the cost function. As our population consists of natural language sentences, the initial individuals must be manually defined for the specific domain. This approach was also adopted in [Guo et al. 2024]. In this work, we base our initial prompts on those proposed in [Oliveira et al. 2023] and [Oliveira et al. 2024], as these studies serve as a baseline and address the same dataset. Subsequently, we derive additional prompts by adding or removing sentences and words, totalizing six prompts. The target model, LLaMA 3.1 8B, performs the classification using only 50 instances from each class to compute the F1-score, given the high computational cost of this function. Following this, prompts are evolved through an iterative process involving two models—the target model LLaMA 3.1 8B and GPT-4o mini over 50 epochs. After this period, the best prompt (with the highest F1-score) is selected and presented below.

“Best Prompt: You are analyzing a social media post. If the text contains hate speech, offenses, aggressions, insults, swear words, or any form of toxicity, respond only with ‘yes’. If it is appropriate and non-toxic, respond ‘no’”.

3.2. Experiment #2 - LLaMA Fine-Tuning Process

To facilitate fine-tuning on modest hardware, we employed a 4-bit quantized version of the LLaMA 3.1 model, using the QLoRA technique [Dettmers et al. 2023], with 8 billion parameters model- LLaMA 3.1 8B [Dubey et al. 2024]². We used the Hugging Face PEFT library³ with the Unsloth library⁴, setting the learning rate to $2e - 4$ and the sequence length to 2048 tokens, while varying the “rank” and “LoRa Alpha” parameters.

The fine-tuning process used the most effective prompt and involved 3,000 training steps, with a batch size of 2 and gradient accumulation set to 4, effectively processing 6000 instances from the training dataset.

Results from experiments varying the parameters “rank” and “LoRa alpha” are presented in Table 1, while the fine-tuning loss function using “rank=16” and “LoRa alpha=16” is shown in Figure 2.

²<https://huggingface.co/unsloth/llama-3-8b-bnb-4bit>

³<https://huggingface.co/docs/peft/index>

⁴<https://github.com/unslothai/unsloth>

3.3. Experiment #3 - GPT-4o mini Fine-Tuning Process

To fine-tune the GPT4-o mini model, we used the Azure AI Studio platform ⁵, leveraging the same training data used in Experiment #2. We adopted the best prompt identified in Experiment #1 and created a JSONL file where each instance of the training set was preceded by the prompt and accompanied by its respective label.

We chose the 2024-07-18 release of GPT4-o-mini, which was the one available for fine-tuning on Azure. After training, the model was deployed on the Azure platform, allowing its use through API calls.

During the evaluation, we noted that utilizing Azure Studio, which incorporates an additional content moderation layer beyond that provided by OpenAI, led to certain moderation inaccuracies. Approximately 1% of the test set was erroneously categorized due to “content moderation errors.” For these instances, we assigned the label “non-toxic.”

3.4. Results Comparison

For comparative purposes, the most effective prompt identified through the iterative prompt evolution approach is tested with three additional models: Maritaca ⁶ AI Sabiá3 [Pires et al. 2023], OpenAI GPT-4o mini [OpenAI et al. 2024] ⁷, and OpenAI ChatGPT 3.5 Turbo [Brown et al. 2020] ⁸, as well as the BERTimbau model [Souza et al. 2020].

The results presented in Table 2 highlight the significance of fine-tuning the LLaMA 3.1 8B model. Specifically, fine-tuning improved the F1-score from 0.61 to 0.75, demonstrating a substantial performance gain. Furthermore, when applying the fine-tuning methodology using the prompt proposed in [Oliveira et al. 2023], the F1-score reached 0.70. However, our prompt evolution approach further improved this to 0.75, indicating that the refined prompt contributed significantly to the model’s performance.

Additionally, the LLaMA 3.1 8B model, despite being fine-tuned with only 3,000 steps and 6,000 instances, performs competitively against other state-of-the-art models like GPT-4o mini, Sabiá3, and BERTimbau. Notably, Sabiá3, a leading model from Maritaca AI, demonstrated comparable accuracy to GPT-4o mini across various high-stakes Brazilian exams, such as OAB, ENEM, and ENADE. These results underscore the effectiveness of our prompt evolution methodology and the potential of smaller models like LLaMA 3.1 8B when paired with efficient fine-tuning techniques.

The results in Table 1 reveal differences in model performance based on the configuration of the “r” (rank) and “LoRa alpha” parameters. The configuration with “r=16” and “alpha=16” achieves the best overall performance, with an F1-Score of 0.75, balancing precision (0.69) and recall (0.83). Increasing “r” to 24 or “alpha” to 24 leads to a marked decline in performance, with the model showing symptoms of overfitting, particularly with a dramatic drop in recall. The configuration with “r=8” and “alpha=16” demonstrates high recall (0.935) but at the cost of precision, indicating a bias towards over-predicting the positive class.

⁵<https://oai.azure.com/portal>

⁶<https://www.maritaca.ai/>

⁷<https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/>

⁸<https://chat.openai.com/>



Figure 2. Loss over steps. Params LoRa alpha = 16 and r = 16.

Table 1. Impact of different r and LoRa alpha configurations on model performance.

Configuration	F1-Score	Precision	Recall	Accuracy
LLaMA 3.1 8B (r=16, alpha=16)	0.75	0.69	0.83	0.76
LLaMA 3.1 8B (r=24, alpha=16)	0.432	0.601	0.338	0.609
LLaMA 3.1 8B (r=8, alpha=16)	0.727	0.595	0.935	0.690
LLaMA 3.1 8B (r=16, alpha=24)	0.327	0.573	0.229	0.584

Table 2. Comparison of Evaluation Metrics for Different Models

Model	F1-Score	Precision	Recall	Accuracy
LLaMA 3.1 8B (original) w/ best prompt	0.61	0.45	0.96	0.46
LLaMA 3.1 8B (finetuned) w/ prompt from [Oliveira et al. 2023]	0.70	0.71	0.70	0.74
LLaMA 3.1 8B (finetuned) w/ best prompt	0.75	0.69	0.83	0.76
ChatGPT 3.5T Zero-Shot w/ prompt from [Oliveira et al. 2023]	0.73	0.74	0.73	0.74
GPT-4o mini w/ best prompt	0.75	0.75	0.75	0.75
GPT-4o mini (finetuned) w/ best prompt	0.74	0.78	0.74	0.76
Sabiá 3 w/ best prompt	0.75	0.77	0.76	0.75
BERTimbau Finetuned	0.75	0.75	0.75	0.75

4. Conclusion

In this study, we investigated whether a smaller, open-source and quantized language model like LLaMA 3.1 8B 4 bits could effectively perform toxic text detection in Portuguese, particularly when optimized using an iterative prompt evolution approach along with finetune. The experiments demonstrated that, with carefully evolved prompts, the model could achieve competitive performance, even with a limited number of training steps and instances. This highlights the potential of smaller models when paired with efficient prompt engineering techniques.

However, the approach has its limitations. The success of the prompt evolution algorithm heavily depends on the quality of the underlying language models used for the text evolution operations. This reliance can be a significant constraint, as deficiencies in the language models directly affect the quality of the evolved prompts and, consequently, the overall model performance. Further research is needed to address these dependencies and enhance the robustness of the prompt engineering approach.

Acknowledgments

We would like to express our sincere thanks to Blip, whose generous support and invaluable assistance were crucial for the presence of the first author in the event. The authors would also like to thank the *Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brazil (CAPES) - Finance Code 001*, *Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG, grants APQ-01518-21)*, *Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq, grants 307151/2022-0, 308400/2022-4)*, and *Universidade Federal de Ouro Preto (PROPPI/UFOP)* for supporting the development of this study.

References

- BehnamGhader, P., Adlakha, V., Mosbach, M., Bahdanau, D., Chapados, N., and Reddy, S. (2024). Llm2vec: Large language models are secretly powerful text encoders.
- Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.
- da Rocha Junqueira, J., Junior, C. L., Silva, F. L. V., Côrrea, U. B., and de Freitas, L. A. (2023). Albertina in action: An investigation of its abilities in aspect extraction, hate speech detection, irony detection, and question-answering. In *Anais do XIV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 146–155. SBC.
- Dettmers, T., Pagnoni, A., Holtzman, A., and Zettlemoyer, L. (2023). Qlora: Efficient finetuning of quantized llms. In Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S., editors, *Advances in Neural Information Processing Systems*, volume 36, pages 10088–10115. Curran Associates, Inc.
- dos Santos, W. R. and Paraboni, I. (2023). Predição de transtorno depressivo em redes sociais: Bert supervisionado ou chatgpt zero-shot? In *Anais do XIV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 11–21. SBC.
- Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Yang, A., and et al., A. F. (2024). The llama 3 herd of models.
- Guo, Q., Wang, R., Guo, J., Li, B., Song, K., Tan, X., Liu, G., Bian, J., and Yang, Y. (2024). Connecting large language models with evolutionary algorithms yields powerful prompt optimizers. In *The Twelfth International Conference on Learning Representations*.
- Hammes, L. O. A. and de Freitas, L. A. (2021). Utilizando bertimbau para a classificação de emoções em português. In *Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana (STIL)*, pages 56–63. SBC.

- Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., Attariyan, M., and Gelly, S. (2019). Parameter-efficient transfer learning for nlp. In International conference on machine learning, pages 2790–2799. PMLR.
- Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. (2021). Lora: Low-rank adaptation of large language models.
- Lee, C., Roy, R., Xu, M., Raiman, J., Shoeybi, M., Catanzaro, B., and Ping, W. (2024). Nv-embed: Improved techniques for training llms as generalist embedding models.
- Lehman, J., Gordon, J., Jain, S., Ndousse, K., Yeh, C., and Stanley, K. O. (2023). Evolution through large models. In Handbook of Evolutionary Machine Learning, pages 331–366. Springer.
- Leite, J. A., Silva, D., Bontcheva, K., and Scarton, C. (2020). Toxic language detection in social media for brazilian portuguese: New dataset and multilingual analysis. In Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing, pages 914–924.
- Meyerson, E., Nelson, M. J., Bradley, H., Gaier, A., Moradi, A., Hoover, A. K., and Lehman, J. (2023). Language model crossover: Variation through few-shot prompting. arXiv preprint arXiv:2302.12170.
- Oliveira, A. S., Cecote, T. C., Alvarenga, J. P. R., Freitas, V. L. S., and Luz, E. J. S. (2024). Toxic speech detection in Portuguese: A comparative study of large language models. In Gamallo, P., Claro, D., Teixeira, A., Real, L., Garcia, M., Oliveira, H. G., and Amaro, R., editors, Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 1, pages 108–116, Santiago de Compostela, Galicia/Spain. Association for Computational Linguistics.
- Oliveira, A. S., Cecote, T. C., Silva, P. H., Gertrudes, J. C., Freitas, V. L., and Luz, E. J. (2023). How good is chatgpt for detecting hate speech in portuguese? In Anais do XIV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana, pages 94–103. SBC.
- OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., and et al., S. A. (2024). Gpt-4 technical report.
- Pereira, P. H. and da Silva, T. L. C. (2023). Uso de modelagem de tópicos para agrupamento de notícias: uma abordagem usando bertopic. In Anais do XIV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana, pages 406–410. SBC.
- Pires, R., Abonizio, H., Almeida, T. S., and Nogueira, R. (2023). Sabiá: Portuguese large language models. In Naldi, M. C. and Bianchi, R. A. C., editors, Intelligent Systems, pages 226–240, Cham. Springer Nature Switzerland.
- Serras, F. and Finger, M. (2021). verbert: Automating brazilian case law document multi-label categorization using bert. In Anais do XIII Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana, pages 237–246, Porto Alegre, RS, Brasil. SBC.
- Souza, F., Nogueira, R., and Lotufo, R. (2020). Bertimbau: pretrained bert models for brazilian portuguese. In Intelligent Systems: 9th Brazilian Conference, BRACIS

2020, Rio Grande, Brazil, October 20–23, 2020, Proceedings, Part I 9, pages 403–417. Springer.

Zheng, M., Su, X., You, S., Wang, F., Qian, C., Xu, C., and Albanie, S. (2023). Can gpt-4 perform neural architecture search? arXiv preprint arXiv:2304.10970.

Disfluency Detection and Removal in Speech Transcriptions via Large Language Models

Pedro L. S. de Lima¹, Cláudio E. C. Campelo¹

¹Departamento de Sistemas e Computação
Universidade Federal de Campina Grande
Campina Grande, Paraíba - Brasil

pedro.lima@ccc.ufcg.edu.br

campelo@dsc.ufcg.edu.br

Abstract. *The field of Automatic Speech Recognition (ASR) has significantly expanded within the technological landscape due to its extensive use in sectors such as education, healthcare, and customer service. Many modern applications depend on analyzing spoken content through Speech-to-Text (STT) conversion models. However, transcriptions produced by these systems often contain undesirable elements, such as word repetitions and the prolongation of certain sounds, known as disfluencies or linguistic crutches. These elements can negatively affect the quality of automatic content analysis by Natural Language Processing (NLP) models, including those for named entity recognition, emotion detection, or sentiment analysis. Therefore, this study aims to evaluate the feasibility of identifying and eliminating linguistic disfluencies using Large Language Models (LLMs), such as GPT-4, LLaMA, Claude, and Gemini, through Prompt Engineering techniques. The approach was tested using a corpus of debate transcriptions with manually annotated disfluency occurrences, yielding promising results.*

1. Introduction

Automatic Speech Recognition (ASR) has become essential in modern society, enabling the conversion of human speech into written text. This technology facilitates a range of applications through Speech-to-Text (STT) models, including virtual assistants, meeting transcription, automatic captioning, and customer service. Despite significant advances in speech recognition accuracy, a constant feature in transcriptions generated by these systems is the presence of linguistic disfluencies. During human speech production, it is common to generate various sounds within speech, known as disfluencies.

Disfluencies have been extensively studied and are primarily classified into three types: hesitations, repetitions, and corrections [Corley and Stewart 2008]. When a speech model transcribes voice into text, it often overlooks the context of the spoken words, focusing instead on achieving an accurate transcription. As a result, these disfluencies are common and appear in caption transcriptions, meeting notes, and any text derived from spontaneous human speech. Various studies explore different techniques for disfluency

detection, ranging from unimodal to multimodal approaches, some even use Transformer-based methods, but none thoroughly investigate the utility of modern and widely accessible Large Language Models (LLMs) for the detection and removal of linguistic crutches.

LLMs based on Transformers present a promising alternative. Due to their ability to capture complex contexts and understand linguistic nuances, such as differentiating between disfluent and fluent text, they present a promising alternative. LLMs can be easily manipulated using Prompt Engineering techniques, which involve creating instructions to guide their behavior toward a specific goal. This work aims to fill a gap in the study of disfluency detection and removal in text transcriptions by evaluating the capabilities of the most advanced LLMs available today, such as OpenAI’s GPT-4o [OpenAI et al. 2024], Gemini 1.5 Pro Experimental 0827 [Team et al. 2024], Claude 3.5 [Anthropic 2024] and LLaMMa 3 (70B parameters) [Meta 2024] to assess the extent of their applicability to this task.

The main contributions of this paper include:

- An analysis of LLMs’ ability to remove particular text excerpts while preserving other relevant information.
- A comparative analysis of available models and their effectiveness in handling transcribed spontaneous human speech.
- An assessment of the feasibility, in terms of computational cost, of cleaning transcriptions of natural human speech.
- A dataset with annotated disfluencies in Brazilian Portuguese.

The following sections of this paper are organized as follows: Section 2 presents a literature review, covering foundational and relevant research on disfluency detection and removal, leading up to the current state-of-the-art. Section 3 details the research methodology, explaining data collection and handling processes, as well as the construction of prompts and an exploratory data analysis, followed by Section 4, which presents the results. Finally, Section 5 offers the conclusion.

2. Related Work

Research on the detection and removal of disfluencies in speech encompasses a variety of techniques, each contributing to the advancement of the state-of-the-art in this field. Studies in this domain typically utilize one of three types of input: textual transcription, audio signal, or a combination of the two. Unimodal solutions rely on a single source of information, whereas multimodal solutions integrate multiple sources, such as audio and text, to perform the task of disfluency detection/removal. The next subsections present research carried out using the unimodal text approach, followed by the unimodal audio approach, a comparison between the two approaches, and finally the conclusion of this section.

2.1. Text-Based Approaches

In this context, [Snover et al. 2004] proposed a Transformation-Based Learning (TBL) algorithm for disfluency detection in speech transcriptions, employing lexical features (word usage and sentence structure). The system, referred to as System A, achieved results comparable to those employing prosodic features (variations in intonation, rhythm,

duration, and intensity/volume of speech), demonstrating that satisfactory performance can be achieved without heavily relying on prosodic cues. The study underscores the importance of features such as the lexeme itself, Part-of-Speech (POS) tags, and word frequency for the speaker in identifying disfluencies. System A showed promising results in detecting various types of disfluencies and paved the way for future research focused on natural language processing techniques.

[Ferguson et al. 2015] proposed a conditional semi-Markovian method (semi-CRF) for disfluency detection in speech transcriptions, focusing on repairs such as repetitions and false starts. This technique utilizes lexical, structural, and prosodic features, such as pauses and word duration, extracted from alignment with the speech signal. This approach achieved an F-score of 85.4% on the Switchboard corpus (a dataset consisting of English telephone conversations collected in the United States during the 1990s), surpassing the performance of previous studies. Concurrently, [Zayats et al. 2016] introduced a novel method for disfluency detection in speech transcriptions using a Bidirectional LSTM (BLSTM) neural network. Their solution employs word embeddings (numerical representations of words), POS tags, and lexical pattern features as input. Additionally, the model incorporates an explicit repair mechanism and uses Integer Linear Programming (ILP) to enforce structural constraints on the disfluency sequence. This approach achieved an F-score of 85.9% on the Switchboard corpus. Analysis of the results indicates that this approach performs better in detecting complex disfluencies that do not involve mere repetitions of words. Despite its effectiveness, the model's reliance on predefined resources limits its adaptability to different types of disfluencies, contexts, and speaking styles.

[Bach and Huang 2019] also explored the BiLSTM technique with self-attention for disfluency detection in speech transcriptions. The authors demonstrated competitive results with BERT on the Switchboard corpus, outperforming it in terms of robustness and efficiency on out-of-domain datasets. The artificial addition of extra and incorrect words during model training proved highly effective in enhancing its robustness to various data types and transcription errors, making it a compelling alternative for disfluency detection in real-world scenarios. Furthermore, the proposed models are smaller than BERT, which results in reduced computational resource requirements overall.

2.2. Audio-Based Approaches

[Bassi et al. 2023] propose an end-to-end approach for speech transcription with disfluency removal using a large-scale pre-trained HuBERT acoustic model. The traditional two-step method, which first transcribes the audio into text and then removes disfluencies, neglects the prosodic cues present in the original audio. The proposed approach processes the audio directly and uses acoustic representations learned during pre-training to identify and remove disfluencies during transcription. The authors demonstrate that the end-to-end solution surpasses the two-step approach in terms of Word Error Rate (WER) and Character Error Rate (CER) on the Switchboard test set, achieving 12.2% WER and 7.3% CER. The study also highlights the significance of the pre-training objective: HuBERT, pre-trained with a clustering objective that groups audio representations based on similarities, significantly outperformed Wav2Vec2, which was pre-trained with a contrastive objective that maximizes similarity among similar samples and minimizes similarity among different samples. These results suggest that end-to-end models with

large-scale acoustic pre-training with clustering objectives are a promising approach for accurate disfluent speech transcription.

2.3. Comparison Between Unimodal and Multimodal Models

[Romana et al. 2023] investigated the automatic detection of disfluencies in speech by comparing language-based, acoustic, and multimodal methods. Their results demonstrated that while language models such as BERT exhibited high accuracy with manual transcriptions, performance significantly declined with the use of transcriptions generated by Automatic Speech Recognition (ASR). Acoustic approaches utilizing models like Wav2Vec 2.0, HuBERT, and WavLM proved promising by avoiding reliance on transcriptions. However, the authors found that multimodal solutions combining acoustic and linguistic information through a BLSTM fusion network achieved the best results, outperforming unimodal techniques in disfluency detection and categorization. This study highlights the potential of multimodal methods for creating more robust disfluency detection systems.

The academic works presented in this section illustrate the progress made in the field, with advanced techniques in artificial intelligence, transformers, and robust multimodal methods applicable to various data types and transcription errors. These solutions have proven effective in detecting and removing disfluencies across diverse contexts. However, the use of widely available Large Language Models (LLMs) for cleaning automatic transcriptions has been insufficiently studied. Therefore, there is a need to investigate how LLMs can be leveraged for this purpose, complementing the advancements achieved in the reviewed academic works and democratizing access to these technologies.

3. Methodology

This section contains information about the methodology used in the research, including how the data was obtained and organized, and the construction of the prompts.

3.1. The Dataset

The dataset for this study consists of text extracted from four debate sessions held at Federal University of Campina Grande to analyze debater performance. It includes 114 minutes of transcribed audio in Portuguese, providing insights into the dynamics and effectiveness of various debating techniques. The debates were moderated, with each session involving 4 to 5 debaters discussing topics related to Artificial Intelligence. Each debater had a chance to speak following questions posed by the moderator, and interruptions were not allowed, resulting in a free-flowing and spontaneous discourse. After the debates, the audio recordings were transcribed using Microsoft's Azure model, with the transcripts stored in a JSON file. This file was then converted into Excel tables containing all the transcribed data. The data underwent human review to correct major transcription errors, such as non-existent or meaningless words. Additionally, each table was annotated for disfluencies. Disfluencies were categorized into three types: hesitations, repetitions, and corrections. Four HTML-style tags were created to mark these disfluencies in the text:

- `<hes {content}/>`, which marks hesitations
- `<rep {content}/>`, which marks repetitions

- `<erro {content}/>`, which marks errors
- `<corr {content}/>`, which marks corrections

This marking and correction process resulted in four Excel files with the transcriptions of the respective debates. These files were then subjected to an exploratory data analysis.

3.2. The Prompts

To perform the task of disfluency detection and removal, four different prompts were developed. To determine which prompt technique is most effective, three types of prompt engineering methods were tested:

- Zero-Shot Prompting
- Few-Shot Prompting
- Chain-of-Thought Prompting

These three types of prompts differ significantly in how they present information to the language model (LLM), and the study aims to understand the extent of the LLMs' knowledge about disfluencies. In the Zero-Shot case, the prompt provides little or no context about the task, so it was divided into two prompts. The first prompt is a direct command to the LLM to remove repetitions, hesitations, and corrections from the text, while keeping it otherwise unchanged. The second prompt adds a description of what disfluencies are and how the three targeted types are characterized. The Few-Shot prompt includes all the information from the first two prompts, as well as an example of disfluent text in three stages: the original disfluent text, the text with disfluency tags, and the cleaned text. Finally, the Chain-of-Thought prompt is designed to help the LLM adopt a step-by-step approach to detecting and removing disfluencies from the text. These four prompts were executed with each of the LLMs. The average number of tokens processed by the LLMs in Group 14, the smallest group, ranged from 4,273 tokens (with the smallest prompt) to 5,041 tokens (with the largest prompt). In contrast, Group 1, the largest group, processed between 5,250 tokens (smallest prompt) and 6,004 tokens (largest prompt). This calculation was estimated using the Tokenizer from the OpenAI Platform.

Prompting Technique	Context
Zero-Shot Prompting	None
Zero-Shot Prompting	Definition of disfluencies
Few-Shot Prompting	Definition of disfluencies and a three-stage snapshot of the text during the disfluency cleaning process
Chain-of-Thought Prompting	Definition of disfluencies plus a guide on how to recognize and remove each type of disfluencies

Table 1. Prompts Created For the Task

3.3. Exploratory Data Analysis

Data from the tagged transcriptions in Excel files were analyzed to gain an overview of how each disfluent text is characterized. The initial analysis focused on the quantity of disfluencies per group. For this purpose, disfluencies were tallied in each file employing the markers described in the Dataset section. These counts were aggregated for each

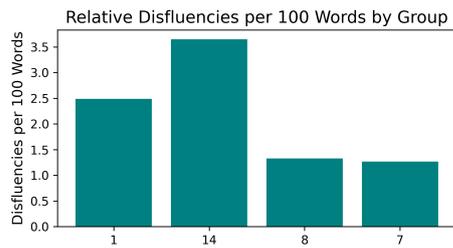


Figure 1. Total Relative Disfluencies per 100 Words

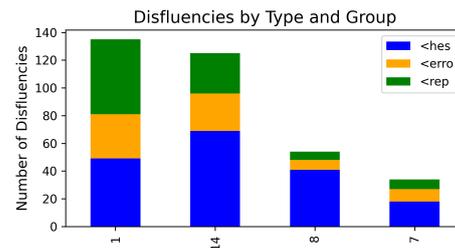


Figure 2. Disfluencies by Type and Group

group, and the totals were visualized using graphs to aid interpretation. Figure 1 displays the comparison of disfluency rates across four groups, labeled 1, 14, 8, and 7 on the X-axis. This figure presents the proportion of disfluencies calculated per 100 words for each group, facilitating a comparison of the relative frequency of disfluencies between the groups. Figure 2, using the same group labels (1, 14, 8, and 7) on the X-axis, depicts the number of disfluencies broken down by type (hesitation, error, repetition). This figure illustrates the distribution of different disfluency types within each group.

3.4. Configuration and Execution of LLMs

The execution of data in Large Language Models (LLMs) was carried out through specific Application Programming Interfaces (APIs). The Google Gemini 1.5 Pro Experimental 0827, Anthropic’s Claude 3.5 Sonnet, and OpenAI’s ChatGPT-4o were accessed via the official APIs provided by their respective companies. The LLaMa 3 72B was used through the Groq platform. The implementation was structured into 16 Python notebooks in the Google Colaboratory environment, with four notebooks assigned to each LLM, corresponding to debate groups. Each notebook was initialized with the configuration of the corresponding LLM, followed by the result extraction codes detailed in this methodological section, and then executed using the pre-established prompts. The results obtained were recorded at the end of each notebook, later compiled into tables for this work, and analyzed for the research objectives. In a separate notebook, an exploratory data analysis was conducted using Excel files from the groups, with the procedures and results described in detail in this work.

4. Results

Table 2. Zero-Shot (No Context) - Group 14

Model	Total Removal Rate	Levenshtein Similarity	Time (seconds)
Gemini	14.06%	97.95%	85.69
GPT-4o	62.50%	96.83%	49.77
LLaMa	53.12%	95.24%	21.86
Claude	57.81%	32.98%	18.47

The data presented in Tables 2 and 3 clearly show that when using Zero-Shot prompts in Group 14 (the most disfluent group), GPT-4o, Gemini, and LLaMa maintained a relatively good textual structure, as indicated by the Levenshtein similarity value. None of the LLMs successfully balanced the removal of disfluencies while preserving the

Table 3. Zero-Shot (With Context) - Group 14

Model	Total Removal Rate	Levenshtein Similarity	Time (seconds)
Gemini	10.94%	97.96%	86.23
GPT-4o	60.94%	74.96%	49.77
LLaMa	62.50%	51.16%	8.79
Claude	60.94%	33.37%	18.65

original text's quality with these two prompts, but text maintenance results for Claude and LLaMa fell significantly below expectations for most tested prompts, making them currently unreliable for this type of task. Therefore, the following analysis focuses solely on GPT-4o and Gemini 1.5.

Table 4. Test Results for GPT-4o and Gemini - Few Shot - Group 14

Model	Total Removal Rate	Levenshtein Similarity	Time (seconds)
Gemini	51.56%	98.18%	82.57
GPT-4o	67.19%	96.83%	74.73

With Few-Shot prompting (Table 4), GPT-4o achieved a 67.19% disfluency removal rate while maintaining 96.83% of the original text. It also surpassed Gemini in response time. Although Gemini had a slightly higher text maintenance rate, it performed poorly in removing disfluencies.

Table 5. Test Results for GPT-4o and Gemini - Chain of Thought - Group 14

Model	Total Removal Rate	Levenshtein Similarity	Time (seconds)
Gemini	26.56%	98.09%	84.61
GPT-4o	68.75%	97.85%	45.76

Using Chain-of-Thought prompting (Table 5), GPT-4o was the only one among the four LLMs to produce a minimally favorable result. When compared to Gemini, GPT-4o achieved a 68.75% total disfluency removal rate, despite a similar text maintenance rate, while Gemini, though maintaining text quality, failed in removing disfluencies.

Table 6. Test Results for GPT-4o and Gemini - Few Shot - Group 8

Model	Total Removal Rate	Levenshtein Similarity	Time (seconds)
Gemini	62.96%	95.80%	102.58
GPT-4o	48.15%	88.22%	47.05

Table 7. Test Results for GPT-4o and Gemini - Chain of Thought - Group 8

Model	Total Removal Rate %	Levenshtein Similarity %	Time (seconds)
Gemini	33.33%	98.19%	104.26
GPT-4o	40.74%	88.51%	45.25

In Group 8, one of the least disfluent groups, GPT-4o's effectiveness declined in both disfluency removal and maintaining text fluency, as shown in Tables 6 and 7.

Table 8. Test Results for GPT-4o and Gemini - Few Shot - Group 1

Model	Total Removal Rate	Levenshtein Similarity	Time (seconds)
Gemini	63.89%	96.68%	126.40
GPT-4o	68.06%	78.18%	52.48

Table 9. GPT-4o and Gemini - Chain of Thought - Group 1

Model	Total Removal Rate	Levenshtein Similarity	Time (seconds)
Gemini	25.00%	97.85%	127.22
GPT-4o	68.06%	77.55%	53.69

Gemini achieved a 62.96% removal rate with good text maintenance, albeit taking more than twice as long. This trend, where GPT-4o did not match Gemini in text maintenance, was also observed in Group 1, as shown in Tables 8 and 9, which is the largest group but not as disfluent as Group 14. The models demonstrated high effectiveness in removing repetitions, achieving 91.30% removal in Group 14 for GPT-4o, compared to Gemini’s 56.52% in the Few-Shot prompt. In the Chain-of-Thought prompt, GPT-4o maintained a consistent removal rate of 91.30% while also outpacing Gemini in processing time. Although GPT-4o showed strong performance in Group 14, it struggled in Group 1, where Gemini achieved 87.18% removal with superior text preservation (96.68%). These results suggest that while GPT-4o excels in specific contexts, Gemini may be more robust when handling larger, more complex texts.¹

5. Conclusion and Future Work

This study explored the efficacy of Large Language Models (LLMs) in detecting and eliminating linguistic disfluencies from transcriptions of academic debates. By leveraging advanced prompt engineering techniques, such as Zero-Shot, Few-Shot, and Chain-of-Thought prompting, we assessed the performance of leading LLMs — GPT-4, Gemini 1.5, Claude 3.5, and LLaMa 3 — in this task. The results revealed several key insights into the capabilities and limitations of these models. GPT-4o demonstrated the highest overall performance in disfluency removal, achieving an optimal balance between removing disfluencies and maintaining text coherence, particularly under Few-Shot and Chain-of-Thought prompting conditions. Gemini 1.5 also performed well but showed variability depending on the prompt type and the specific debate group analyzed. It excelled in text maintenance but had lower removal rates compared to GPT-4o in some cases. Claude 3.5 and LLaMa 3 produced weaker results, struggling to maintain text coherence while removing disfluencies. GPT-4o demonstrated more efficient processing times compared to the other models, which is crucial for practical, real-world applications. In conclusion, while LLMs like GPT-4o and Gemini 1.5 show promise for improving transcription quality by removing disfluencies, further advancements—such as fine-tuning, employing more advanced prompt engineering techniques, integrating widely used LLMs with multimodal systems, or developing future models—are necessary to fully enhance their capabilities.

¹Repository: <https://github.com/pedrosqra/STIL>

References

- Anthropic (2024). Claude 3.5 sonnet. <https://www.anthropic.com/news/claude-3-5-sonnet>. Accessed: 2024-08-27.
- Bach, N. and Huang, F. (2019). Noisy bilstm-based models for disfluency detection. In *Interspeech*, pages 4230–4234.
- Baevski, A., Zhou, H., Mohamed, A., and Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. *CoRR*, abs/2006.11477.
- Bassi, S., Duregon, G., Jalagam, S., and Roth, D. (2023). End-to-end speech recognition and disfluency removal with acoustic language model pretraining. *arXiv preprint arXiv:2309.04516*.
- Corley, M. and Stewart, O. W. (2008). Hesitation disfluencies in spontaneous speech: The meaning of um. *Language and Linguistics Compass*, 2(4):589–602.
- Ferguson, J., Durrett, G., and Klein, D. (2015). Disfluency detection with a semi-markov model and prosodic features. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 257–262.
- Hsu, W., Bolte, B., Tsai, Y. H., Lakhotia, K., Salakhutdinov, R., and Mohamed, A. (2021). Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *CoRR*, abs/2106.07447.
- Meta, A. (2024). Introducing llama 3: Advancements in large language models. Accessed: 2024-08-27.
- OpenAI (2024). Openai tokenizer. Accessed: 2024-10-08.
- OpenAI, Achiam, J., and et al., S. A. (2024). Gpt-4 technical report.
- Romana, A., Koishida, K., and Provost, E. M. (2023). Automatic disfluency detection from untranscribed speech. *arXiv preprint arXiv:2311.00867*.
- Snover, M., Dorr, B., and Schwartz, R. (2004). A lexically-driven algorithm for disfluency detection. In *Proceedings of HLT-NAACL 2004: Short Papers*, pages 157–160.
- Team, G., Georgiev, P., and et al., V. I. L. (2024). Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context.
- Zayats, V., Ostendorf, M., and Hajishirzi, H. (2016). Disfluency detection using a bidirectional lstm. *arXiv preprint arXiv:1604.03209*.

Detection and Censorship of Offensive Language in Extended Texts in Portuguese

Lucas Lenocho de Souza^{1,2}, Franciele Beal^{1,3},
André Roberto Ortoncelli^{1,2,4}, Marlon Marcon^{1,2,4}

¹Federal University of Technology - Paraná - UTFPR

²Software Engineering Coordination - Dois Vizinhos - Paraná - Brazil

³Academic Department of Informatics - Pato Branco - Paraná - Brazil

⁴Programa de Pós-graduação em Informática (PPGI)

lucasouza191141@gmail.com

{fbeat, ortoncelli, marlonmarcon}@utfpr.edu.br

Abstract. *This article addresses the problem of detecting and censoring offensive language in extensive Brazilian Portuguese texts on the web. This paper proposes a pipeline for classifying and censoring extensive texts, focusing on comments, posts, and articles using NLP techniques. The results include an in-depth review of current methods for offensive content classification in Portuguese and the implementation of a BERTimbau-based pipeline for offense detection. This work represents a significant advancement in the state-of-the-art NLP in Portuguese, promoting safer and more respectful online environments for users, especially children.*

1. Introduction

In recent years, the Internet has been growing at an impressive rate in terms of users and the data generated by web page publications. This increasing Internet use often introduces many children to virtual environments from a very early age. Consequently, offensive language in online texts becomes a concern for ethical reasons [Economist 2019]. Issues such as cyberbullying, hate speech, and various forms of offensive content in social media posts are also relevant [Cook 2024].

Regarding the intelligent interpretation of textual data from social networks, Natural Language Processing (NLP) is commonly used. To address the problem of offensive language in web posts, articles in the field of NLP, combined with Machine Learning (ML) and Deep Learning (DL) techniques, have been developed [Hajibabae et al. 2022]. These efforts include creating research pipelines and developing high-quality annotated datasets by professionals [Leite et al. 2020].

Given the differences in languages and cultures, it is only possible to formalize a model for some languages. However, efforts are being made to learn various languages [Husain and Uzuner 2021]. The HateBR [Vargas et al. 2022] corpus is an example for Brazilian Portuguese. Such an article highlights the lack of academic production related to offensive language in the national language, along with the dataset and its results. Despite existing academic contributions to addressing offensive language in Brazilian

Portuguese, applying NLP and ML/DL techniques to classify longer texts (such as news articles or blog posts) remains challenging. These techniques are often limited to social media posts or news comments.

Based on this information, the present work focuses on detecting and subsequently censoring or filtering offensive language online. The goal is to ensure that web pages that are not explicitly focused on adult content can be suitable environments for children. Additionally, efforts are made to reduce prejudice and offensiveness in online posts, benefiting users' emotional well-being and promoting mutual respect. Specifically, we propose identifying offensive words using DL techniques, which can then be filtered or censored. We define small texts as comments, posts, and sentences, while paragraphs and entire pages constitute more extended texts, with the latter being the focus of this study.

As the main contributions of this work, we have: 1) An in-depth review of state-of-the-art methods applied to offensive content classification for the Portuguese language; 2) An update to the state-of-the-art results on the HateBR dataset [Vargas et al. 2022]; and 3) A Deep-Learning-based pipeline that effectively classifies and censors extensive Brazilian Portuguese texts based on their offensiveness.

2. Related Works

When considering works that address text processing with offensive language in the Brazilian context, the range of existing works is very small, including ToLD-BR [Leite et al. 2020], OffComBr [Pelle and Moreira 2017], HateBR [Vargas et al. 2022], and OLID-BR [Trajano et al. 2023]. Such methods are explained in the following sections.

2.1. ToLD-BR

This dataset [Leite et al. 2020], presents an unspecified number of posts extracted from the Twitter platform. A total of 42 individuals, chosen from 129 volunteers, were tasked with annotating each post, classifying them into various categories of prejudice: homophobia, obscene language, misogyny, and xenophobia.

The posts were classified using BERT-style algorithms [Devlin 2018], which achieved optimal results for such a complex and subjective task. The researchers also explored the possibility of building models for this task in multiple languages, but their results indicated that monolingual data is still preferable for more accurate classifications.

2.2. OffComBr

This paper presents the development of a dataset comprising comments on news articles derived from the website g1.globo.com, named "OffComBr" [Pelle and Moreira 2017]. The researchers obtained around 10,000 comments, but given the manual annotation process done by three experts in detecting offensive language, they included only 1,250 comments in the final dataset.

The authors performed two classification algorithms (SMO and Naive Bayes) to evaluate the dataset, with different assessments depending on different data preprocessing methods. Two versions of the dataset were developed, OffComBR-2 and OffComBR-3, with the difference being the size, as the latter retained from the former only the annotations agreed upon by all three experts.

2.3. HateBR

The work of [Vargas et al. 2022] presented the first large annotated corpus of offensive language in Instagram comments in Brazilian Portuguese. Motivated by the presence of hate speech on social media and the lack of studies on the subject in Portuguese, the project collected 7,000 Instagram comments, annotated by experts regarding the presence, degree, and category of offensiveness. The process involved data collection, selection of accounts of Brazilian political figures (three left-wing and three right-wing), and the selection of 30 posts from which 15,000 comments were extracted, with 7,000 being balanced between offensive and non-offensive.

The comments were labeled into three levels of offensiveness: whether they were offensive or not, the degree of offensiveness (mild, moderate, or high), and whether they contained hate speech, categorized into nine types such as xenophobia, racism, and homophobia. From the 7,000 comments, 3,500 were offensive, with 778 highly offensive, 1,044 moderately offensive, and 1,678 mildly offensive. Among the offensive comments, 727 contained some type of hate speech. Table 1 presents samples of data from HateBR.

Table 1. Examples of comments extracted from the HateBR dataset.

Class	Comments
Offensive	Essa besta humana é o câncer do País, tem que voltar para a jaula, urgentemente! E viva o Presidente Bolsonaro.
Non-Offensive	Quem falou isso para você deputada? O Sergio Moro está aprovado pela maioria dos brasileiros.
With hate speech	Vagabunda. Comunista. Mentirosa. O povo chileno não merece uma desgraça dessa.
Without hate speech	Pois é, deveria devolver o dinheiro aos cofres públicos do Brasil. Canalha.

Finally, after a detailed explanation of their entire annotation system, as well as evaluations to judge the annotations of each of the three experts and decide the most appropriate annotations for each comment, the study presents the test results with some ML models trained on the HateBR corpus, comparing the best result obtained with the best results of two other reference works. In this work, we seek to replicate the results obtained by HateBR, following the same training procedure and using the same models for comparison with our trained model.

2.4. OLID-BR

The work of [Trajano et al. 2023] also developed an annotated dataset of offensive comments in Portuguese, similar to HateBR. However, the main advantage of this dataset lies in its application to various NLP tasks, including binary classification of offensiveness, multi-category prediction of the type of toxicity, identification of targeted toxic comments, prediction of the target of toxicity, and identification of toxicity spans in comments.

The primary focus of the work was on the task of identifying toxicity spans, which involves detecting sequences of characters containing offensive language. To collect data, OLID-BR used various sources such as Twitter, YouTube, and other datasets with different annotation schemes.

The annotation was conducted in three stages: detection of offensive language, categorization of offensive language, and identification of the target of the offense. Com-

pared to HateBR, OLID-BR distinguishes between offensiveness against an individual, a group, or another type of target, while HateBR focuses on categorizing hate speech.

Data annotation in OLID-BR was not exclusively done by humans but also with the assistance of the Perspective API¹, allowing human annotators to correct the classifications. The entire corpus was divided into three datasets for training and testing, with a similar distribution of classifications in each. This work replicated the part of OLID-BR related to the identification of toxicity spans, using the code available on GitHub to train the model and apply it to the HateBR dataset for detecting offensive phrases and to the OLID-BR for identifying offensive spans.

3. Main Technologies

For the development, training, and testing of techniques for offensive language detection and censorship, we primarily used two libraries available for the Python language for developing ML algorithms: Transformers and spaCy.

3.1. Transformers

The Transformers library is a Python tool that offers state-of-the-art architectures for NLP tasks, featuring over 32 pre-trained models in more than 100 languages. It provides deep interoperability between TensorFlow 2.0 and PyTorch. The library is named after the Transformer architecture introduced by Google Brain in 2017, which is based on the "attention mechanism." This mechanism allows the model to focus on important parts of the input data, leading to superior performance in NLP tasks like sentence classification, named entity recognition (NER), and natural language generation compared to previous models like recurrent neural networks (RNNs) [Vaswani et al. 2017].

3.2. BERT and BERTimbau

The algorithm used in the first stage of our pipeline was a fine-tuned version BERTimbau [Souza et al. 2020], a Brazilian model based on BERT (Bidirectional Encoder Representations from Transformers) [Devlin 2018]. BERT, introduced by Google in 2018, is a language model that generates numerical representations for words based on their surrounding context and is used for various NLP tasks. BERTimbau adapts BERT for Brazilian Portuguese using transfer learning, where a BERT model was trained on a Portuguese corpus (brWaC) [Wagner Filho et al. 2018] and evaluated on tasks like sentence similarity, textual entailment, and named entity recognition. In this work, BERTimbau was specifically used to develop a model for detecting offensive language in sentences.

3.3. SpaCy

SpaCy is an open-source NLP library for Python, written in Cython [Honnibal et al. 2020], that facilitates tasks like part-of-speech tagging, named entity recognition (NER), and dependency parsing. It provides pre-trained models and allows users to train their own models for NER, where sentences are segmented into words, each categorized (e.g., nouns, adverbs, or specific problem-related categories like offensive and non-offensive). In this work, SpaCy was used in the second stage to detect which words in an offensive sentence are offensive, following the methodology of OLID-BR, which also used SpaCy for tasks like detecting offensive spans in text.

¹<https://www.perspectiveapi.com/>

4. Methodology

This study focused on developing NLP models to detect offensive language in extended texts. Both the acquisition of training data and the actual censorship of words detected as offensive were carried out simplified due to them not being the primary focus. The development process consisted of four stages: data collection, cleaning/tokenization, model training, and result evaluation explained in the following subsection.

4.1. Data Collection

For data collection, we employ the HateBR [Vargas et al. 2022] and OLID-BR [Trajano et al. 2023] datasets to train the offensive content detection models, the former to classify a text segment as potentially offensive or not and the latter to identify words or expressions that contain offensiveness. Additionally, to evaluate qualitatively our solution, we selected news articles from the G1 portal [Monteiro 2023], *Catraca Livre* [Leray 2023], and two blog posts from *Senso Incomum* [Trielli 2021, Martins 2022].

4.2. Data Cleaning and Feature Extraction

We used tools from the Transformers and spaCy libraries for data cleaning and feature extraction. Specifically, for the model trained with the Transformers library, we employed a tokenizer ready to transform sentences into numerical representations used by the model for calculations. For the model trained with spaCy, an embedded tokenization functionality was available. In summary, in this step, the trained models were capable of cleaning the data and processing it without needing external code. For the training and evaluation of the first model (developed with the Transformers library), we compared it to the ML models presented by HateBR, which had the best results, using the same feature extraction method they did: TF-IDF. TF-IDF (Term Frequency–Inverse Document Frequency) calculates how relevant a word in a corpus is to a text, obtained by the ratio between the number of times the term in question appears in one of the corpus texts (Term Frequency) and the frequency of appearances of this same term in the entire corpus (Inverse Document Frequency). The frequency of the word in a text refers to the ratio between the number of times it appears in the text and the number of words in the text, while the frequency of the word in the corpus is the count of how many times it appears in all texts of the dataset. The reason for using such feature extraction is due to the empirical results demonstrated by HateBR [Vargas et al. 2022], which show that, in general, models trained with features extracted using TF-IDF outperformed other methods.

4.3. Model Training

The model training process was split into two parts: the first was responsible for detecting offensiveness in a text (in this case, in a selected paragraph), and the second was responsible for identifying words or expressions that contain offensive language in segments classified as offensive by the first model.

We employed the BERTimbau model from the HuggingFace platform for the first model. BERTimbau is a Brazilian language model developed by NeuralMind [Souza et al. 2020] through a technique called fine-tuning. In the case of this study, this process involved adding an extra layer of neurons at the end of the model, which was responsible for classifying an input text as offensive or non-offensive.

For the second model, the training process used by the OLID-BR study [Trajano et al. 2023] was utilized to detect offensive spans (i.e., sequences of characters containing offensive language, not limited to isolated words but also including expressions and punctuation).

4.4. Model Demonstration

We evaluate the offensive language detection models using Precision, Recall, and F1-Score metrics. Precision measures how much we can trust a model when it predicts that an example belongs to a particular class by calculating the number of examples the model correctly predicted as belonging to that class divided by the total number of examples it predicted as belonging to that class. Recall is the number of samples the model correctly identified as belonging to a class divided by the total number of samples that belong to that class in the data. F1-Score is the harmonic mean between precision and recall, i.e., it is the average of both precision and recall values, giving more importance to low values, as a much lower precision or recall value indicates that the model is not balancing these two metrics well when we want to give equal importance to both.

To demonstrate the models' effectiveness, we conducted qualitative tests on selected texts, as long as they were at least one page long or had more than one paragraph.

For the application process of the trained models, Figure 1 presents the following steps graphically: (1) a large text is collected; (2) the text is divided into fragments based on the characters of periods, commas, semicolons, exclamations, questions, and new lines; (3) for each fragment, the BERTimbau-based model is used to check if it is offensive; (4) if not offensive, the fragment is returned as usual, but if it is, it proceeds to the next step; (5) the spaCy-trained model is used to identify the offensive spans; (6) returning of the offensive censored spans, and the censored version of the fragment; (7) finally, all fragments, censored or not, are reassembled using the same separators as before, thus returning the entire text now with the appropriate censorship.

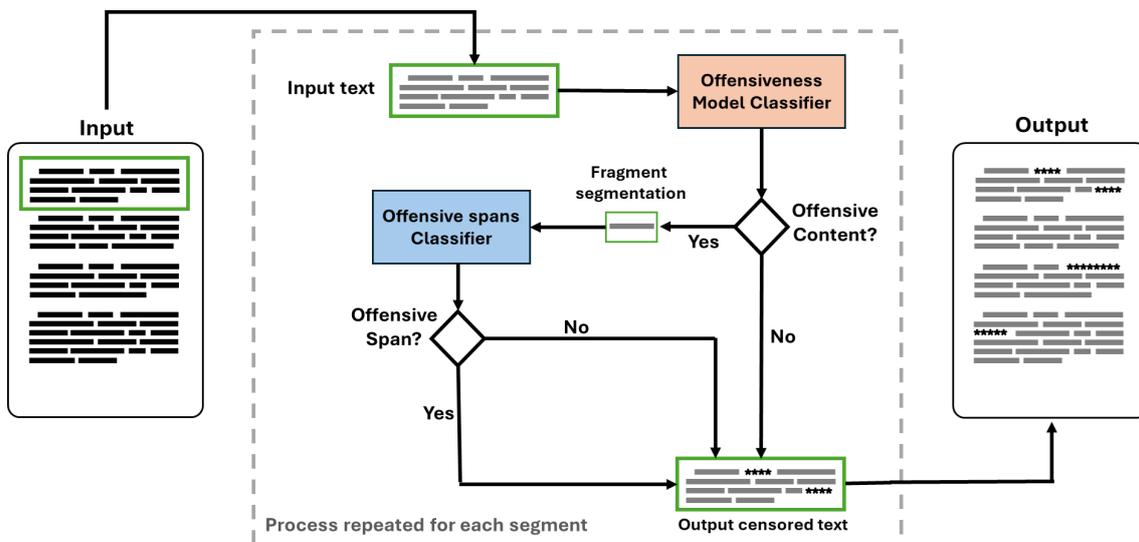


Figure 1. Censorship process of our proposal.

5. Results

The results obtained from the training sessions, in terms of Precision, Recall, and F1-score evaluated across the four different models tested for toxicity detection in sentences, are presented in Table 2. The parameters for conducting the training and testing followed the proposal in the HateBR study [Vargas et al. 2022]. Since HateBR does not provide the original pre-trained models and also the train/validation/test split configuration, it was necessary to retrain each of the four models on this dataset to guarantee consistency between the results: SVM, Naive Bayes, Logistic Regression, and MLP models. We also evaluated our fine-tuned version of the BERTimbau model [Souza et al. 2020], following the TF-IDF method for feature extraction. For training the BERTimbau-based model, we utilized the tokenizer that comes pre-built with the model to perform feature extraction, as this is how BERT-based models operate.

Model	Precision	Recall	F1-score
SVM	0.87	0.84	0.86
NaiveBayes	0.85	0.87	0.86
LogisticRegression	0.86	0.85	0.85
MLP	0.86	0.82	0.84
BERTimbau + <i>fine-tuning</i>	0.92	0.87	0.90

Table 2. Comparison of trained models on the HateBR dataset [Vargas et al. 2022].

The data analysis shows that the BERTimbau + *fine-tuning* model employed in this study outperforms previous results in all comparison parameters, establishing this work as the state-of-the-art for the HateBR dataset. This result underscores the importance of using DL models in the NLP context, as they can yield significant benefits in text recognition and classification processes.

Having completed the test of the BERTimbau model and trained the spaCy model following the same methodology as OLID-BR [Trajano et al. 2023], we conducted the final evaluation of the complete pipeline. For this purpose, we executed the pipeline proposed in Figure 1, i.e., if a sentence is toxic, the offensive parts are detected and censored. For presentation and qualitative evaluation purposes, censorship is performed by simply replacing the characters that constitute the offensive word or expression with asterisks (*).

We selected some news articles from specialized portals, such as G1, Estadão, and Catraca Livre, to perform a preliminary test, but no offensive language was detected. Subsequently, an opinion article about Chico Buarque from the Senso Incomum blog [Martins 2022] was used, where the presence of swear words, which would be appropriately detected, was evident.

Table 3 presents examples for qualitative analysis of the results of applying the proposed pipeline in this work. The Table shows the comparison between the original text and the censored text. The demonstrated text is the only post from a collected blog, as the three obtained news articles contained no offensive language. These results are limited to demonstrating the parts of the original text that our algorithm censored.

One consideration is that the model censored the word “*censura*” (in english censorship), which is usually not considered offensive, as well as the expression “*Rock das*

Original Text Part	Censored Part
observador de bonobos	observador de *****
A autocensura é o pior tipo de censura que existe	A auto***** é o pior tipo de ***** que existe
Joga pedra na Geni/Joga bosta na Geni/Ela é feita pra apanhar/Ela é boa de cuspir /Ela dá pra qualquer um/Maldita Geni!	Joga pedra na Geni/Joga ***** na Geni/Ela é feita pra *****Ela dá pra qualquer um/Maldita Geni
seu lesbofóbico “Rock das Aranhas”!	seu lesbofóbico *****
Paga-pau dos porcos estadunidenses , com toda certeza!	*****, com toda certeza!
Espumando de ódio	Espumando de ****

Table 3. Examples of original text excerpts (on the left) and censored excerpts (on the right).

Aranhas” (Rock of the Spiders) instead of the previous word “*lesbofóbico*” (lesbophobic), which we judged to be far more offensive than an expression about rock and spiders. The remaining censorship we considered appropriate.

5.1. Source Code of the Experiments

The source code developed has been made publicly available in the form of Notebooks, available on the GitHub Repository: <https://github.com/ICDI/censorship-offensive-language>:

- The notebook for training and testing the model based on the BERTimbau model;
- The notebook for training the spam detection model with spaCy, reusing the code from OLID-BR
- The notebook with the tests using both models for text detection and censorship.

6. Conclusion

This work developed an ML/DL-based program for detecting and censoring offensive language in extended Portuguese texts extracted from the web. To this end, we present a pipeline comprising two parts: one that detects the offensiveness in a portion of text (precisely a sentence) and another that detects the offensive parts (spam) within the sentence. This allows for censoring a large text part by part and returning the censored text.

A positive aspect of this work is providing a program specifically focused on detecting and censoring large texts, presenting satisfactory results in quantitative and qualitative analyses. As an evolution of this work, developing a specific dataset for large text censorship can be envisaged, which, unlike the datasets used [Vargas et al. 2022] and [Trajano et al. 2023], would represent a significant advancement.

The results for detecting offensiveness in texts were superior to our original reference (the HateBR article), representing a new benchmark concerning the state-of-the-art, surpassing the techniques used as a reference in the HateBR dataset [Vargas et al. 2022]. The censorship part was performed simply by replacing the characters in the offensive parts with asterisks, which can undoubtedly be improved by future work. The analysis could have been more robust due to the lack of a specific dataset for large texts, even when discussing qualitative data. This fact demonstrates the need for creating a specific dataset built from expert judgments on offensive language or by calculable metrics, which can evolve in future works.

References

- Cook, S. (2024). Cyberbullying statistics and facts for 2024 | comparitech. <https://www.comparitech.com/internet-providers/cyberbullying-statistics/>. (Accessed on 10/10/2024).
- Devlin, J. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Economist (2019). Increasing numbers of children have internet addiction – how worried should parents really be? <https://inews.co.uk/news/long-reads/internet-addiction-children-increase-parents-guide-242434>. (Accessed on 10/10/2024).
- Hajibabae, P., Malekzadeh, M., Ahmadi, M., Heidari, M., Esmailzadeh, A., Abdolazimi, R., and Jones, J. H. (2022). Offensive language detection on social media based on text classification. *2022 IEEE 12th Annual Computing and Communication Workshop and Conference (CCWC)*, pages 0092–0098.
- Honnibal, M., Montani, I., Van Landeghem, S., and Boyd, A. (2020). spaCy: Industrial-strength Natural Language Processing in Python.
- Husain, F. and Uzuner, O. (2021). A survey of offensive language detection for the arabic language. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 20(1):1–44.
- Leite, J. A., Silva, D., Bontcheva, K., and Scarton, C. (2020). Toxic language detection in social media for brazilian portuguese: New dataset and multilingual analysis. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 914–924.
- Leray, W. (2023). Série de harry potter? envolvimento de j.k. rowling divide fãs. <https://catracalivre.com.br/entretenimento/nova-serie-de-harry-potter-polemica-envolvendo-j-k-rowling-divide-fas/>. (Accessed on 09/02/2024).
- Martins, T. (2022). Chico buarque dá comida aos censores - senso comum. <https://sensoincomum.org/2022/01/28/chico-buarque-da-comida-aos-censores/>. (Accessed on 09/02/2024).
- Monteiro, E. (2023). Caso bruno e dom: justiça decide levar amarildo e outros dois réus a júri popular | amazonas | g1. <https://g1.globo.com/am/amazonas/noticia/2023/10/03/caso-bruno-e-dom-justica-decide-levar-amarildo-e-outros-dois-reus-a-juri-popular.ghtml>. (Accessed on 09/02/2024).
- Pelle, R. P. and Moreira, V. P. (2017). Offensive comments in the brazilian web: a dataset and baseline results. In *Anais do VI Brazilian Workshop on Social Network Analysis and Mining*. SBC.
- Souza, F., Nogueira, R., and Lotufo, R. (2020). Bertimbau: pretrained bert models for brazilian portuguese. In *Intelligent Systems: 9th Brazilian Conference, BRACIS 2020, Rio Grande, Brazil, October 20–23, 2020, Proceedings, Part I* 9, pages 403–417. Springer.

- Trajano, D., Bordini, R. H., and Vieira, R. (2023). Olid-br: offensive language identification dataset for brazilian portuguese. *Language Resources and Evaluation*, pages 1–27.
- Trielli, L. (2021). Escócia: estupradores que se declararem mulher serão colocados em prisões femininas - senso incomum. <https://sensoincomum.org/2021/12/14/escocia-estupradores-que-se-declararem-mulher-serao-colocados-em-prisoas-femininas/>. (Accessed on 09/02/2024).
- Vargas, F., Carvalho, I., Rodrigues de Góes, F., Pardo, T., and Benevenuto, F. (2022). HateBR: A large expert annotated corpus of Brazilian Instagram comments for offensive language and hate speech detection. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7174–7183, Marseille, France. European Language Resources Association.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.
- Wagner Filho, J. A., Wilkens, R., Idiart, M., and Villavicencio, A. (2018). The brwac corpus: a new open resource for brazilian portuguese. In *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*.

EyetrackingMOS: Proposta de um método de avaliação online para modelos de síntese de fala

Gustavo E. Araújo¹, Julio C. Galdino¹, Rodrigo de F. Lima¹, Leonardo Ishida¹, Gustavo W. Lopes¹, Miguel Oliveira Jr.², Arnaldo Candido Jr.³, Sandra M. Aluísio¹, Moacir A. Ponti¹

¹ Universidade de São Paulo (USP)

² Universidade Federal de Alagoas (UFAL)

³ Universidade Estadual de São Paulo (UNESP)

{gustavo.evangelista, juliogaldino, guico21, leoishida, gustavowlopes}@usp.br
miguel@fale.ufal.br, arnaldo.candido@unesp, {moacir, sandra}@icmc.usp.br

Abstract. *Evaluating Text-To-Speech (TTS) systems is challenging, as the increasing quality of synthesis makes it difficult to discriminate models' ability to reproduce prosodic attributes, especially for Brazilian Portuguese. Offline evaluation metrics do not capture our genuine reactions to audio stimuli. Therefore, we propose an online evaluation method using eye-tracking. Our experiments with 76 annotators show a reasonable correlation between EyetrackingMOS and MOS, as well as a reduction in the total evaluation time. We believe this metric provides precise and potentially fast information to complement existing evaluation methods.*

Resumo. *Avaliar sistemas Text-To-Speech (TTS) é um desafio, uma vez que a qualidade crescente da síntese impõe obstáculos em discriminar a capacidade de modelos em reproduzir atributos prosódicos, especialmente para o português brasileiro. Métricas de avaliação offline não medem a reação genuína de avaliadores aos estímulos de áudios. Propõe-se, portanto, um método de avaliação online com rastreamento de globo ocular. Os experimentos com 76 anotadores apontam que há uma correlação razoável entre EyetrackingMOS e MOS, assim como uma redução em sua duração total. Desta forma, acredita-se que esta métrica forneça uma informação precisa e potencialmente rápida para complementar os métodos de avaliação.*

Index Terms: Speech Synthesis Models Evaluation, Portuguese language, spontaneous speech, eyetracking

1. Introdução

Sistemas de texto-para-fala, do inglês *Text-To-Speech* (TTS) buscam vocalizar um texto escrito em níveis próximos a naturalidade de fala humana [Caseli and Nunes 2024]. Os avanços em Aprendizado Profundo impulsionaram o desenvolvimento de tais sistemas. Posteriormente, a utilização de modelos gerativos baseados em fluxo, como os propostos por [Kingma et al. 2016] e [Hoogeboom et al. 2019], tem permitido maior flexibilidade

na manipulação de características prosódicas¹ da fala sintética. Os resultados de modelos do estado da arte já reproduzem a identidade dos locutores com bastante naturalidade em condições mais amplas de dados [Casanova et al. 2022, Tan et al. 2022].

Entretanto, modelos de síntese ainda encontram obstáculos na reprodução de aspectos específicos da expressividade individual de falantes. Estes aspectos podem ser medidos através da entoação, duração e ritmo da fala [Ju et al. 2024], que são de natureza prosódica, o que se agrava em cenários de síntese *zero-shot* [Casanova et al. 2022, Ju et al. 2024]. Neste contexto, sistemas contemporâneos de TTS investigam outras capacidades além da reprodução da identidade de um locutor com naturalidade nos resultados, dentre elas o interesse em manter a naturalidade ao gerar fala nas variantes internacionais de uma língua (**accent-robust**), como o *Synthesizing Multi-Accent Speech By Weight Factorization* (SYNTACC) [Nguyen et al. 2023]. A possibilidade de síntese de fala com sensibilidade de sotaques internacionais, também levanta hipóteses de aplicações para variantes linguísticas regionais de uma dada língua que conta com menos recursos, afim de avaliar se a qualidade se preserva. O português brasileiro é uma língua que contempla uma grande quantidade de variantes, dadas as dimensões continentais do Brasil, e devido a fatores históricos, sociais e culturais [Mota et al. 2023].

Para avaliar a qualidade da fala sintetizada nesses sistemas, são utilizadas diversas métricas. As **métricas subjetivas** como: *Mean Opinion Score* (MOS) [ITU - T 1996], *Crowd MOS* [Ribeiro et al. 2011], *Similarity MOS* (SMOS) [Jia et al. 2019] e *Comparative MOS* (CMOS), por um lado, dependem da opinião e percepção de um grupo de ouvintes humanos. Apesar de importante, este perfil de métricas pode oferecer risco para análise de sotaques a depender da correspondência entre o contexto regional/cultural dos avaliadores e os áudios sintéticos, uma vez que a avaliação será influenciada por seus contextos culturais, linguísticos e experiências individuais. Por outro lado, as **métricas objetivas** como: *Speaker Encoder Cosine Similarity* (SECS) [Casanova et al. 2021], *Prosody Similarity with Prompt*, *Prosody Similarity with Ground Truth* e *Word Error Rate* (WER) [Shen et al. 2023] podem não capturar completamente a percepção humana da qualidade do áudio sobre o desempenho na qualidade de expressividade individual e representatividade de variantes linguísticas e, por isso, complementam a análise subjetiva. A ausência de uma métrica padrão e amplamente aceita dificulta a identificação de tendências e avanços consistentes no campo do TTS, além de dificultar o entendimento de quais modelos são mais adequados para determinados cenários ou requisitos específicos (cf. [Le Maguer et al. 2024]). Ambos os perfis têm sensibilidades a aspectos diferentes e limitações que devem ser avaliadas [Cooper et al. 2024].

Ambas as métricas também podem ser observadas quanto a sua resposta aos estímulos de áudios fornecidos durante a avaliação. Em métodos de **avaliação offline** (MOS, CrowdMOS, SMOS e CMOS), o indivíduo pontua apenas após ouvir todo o estímulo, enquanto métodos de **avaliação online** permitem que se registre suas impressões à medida que o estímulo é recebido, tendo como objetivo capturar reações genuínas e momentâneas. A avaliação de estímulos de áudio utilizando rastreamento ocular já é amplamente empregada em contextos linguísticos, como na

¹A prosódia estuda as funções dos suprasegmentos, que são essenciais para a melodia da fala (tom, entoação, tessitura), para a dinâmica da fala (duração, pausa etc.) e para qualidade da voz (volume, registro etc.) [Cagliari 1992].

análise de processamento de linguagem, compreensão auditiva, e percepção fonética [ALMEIDA et al. 2021]. No entanto, sua aplicação na avaliação de sistemas de síntese de fala ainda é pouco explorada. Buscamos preencher essa lacuna, propondo um novo método, EyetrackingMOS, que utiliza o rastreamento ocular para avaliação de qualidade dos áudios forma mais natural, sem que o participante atribua uma nota de forma direta.

As principais contribuições feitas nesse trabalho são sumarizadas como se segue:

1. Proposta de um novo método de avaliação de sistemas de síntese de fala que integra o rastreamento ocular, chamado de EyetrackingMOS;
2. Comparação entre o EyetrackingMOS e uma adaptação do MOS tradicional, destacando suas respectivas vantagens e limitações;
3. Apresentação dos experimentos, detalhes de configuração do modelo e interfaces em um repositório², facilitando a replicabilidade em diferentes cenários e promovendo avanços na pesquisa sobre síntese de fala.

2. Revisão sobre métricas subjetivas para análise de sistemas de TTS

Na década de 1990, a *International Telecommunication Union* (ITU) padronizou diversos tipos de testes de audição que eram frequentemente usados na telefonia [ITU - T 1996]. A pontuação baseada em opinião pode ser definida como o valor em uma escala predefinida que um sujeito atribui à sua opinião sobre o desempenho de um sistema [ITU - R 2017, Loizou 2011]. A pontuação média de opinião, do inglês *Mean Opinion Score* (MOS) é um tipo de *Absolute Category Rating* (ACR) [Ribeiro et al. 2011]. A MOS emergiu como o descritor mais popular sobre a percepção da qualidade de mídia. Para o cálculo da MOS, humanos avaliam os áudios sintetizados e naturais e atribuem uma nota de 1 a 5, no qual o valor final corresponde à média das notas de todos os avaliadores. A tabela traduzida com a equação correspondente pode ser vista no repositório².

Diversas variações da MOS foram desenvolvidas para atender a diferentes necessidades de avaliação. A *Crowd Mean Opinion Score* (crowdMOS) propõe uma adaptação ao ambiente tradicional de testes MOS, ao utilizar trabalhadores de uma multidão (do inglês, *crowd*) pela internet para realizar avaliações em ambientes não controlados, o que permite maior diversidade de ouvintes a um custo reduzido, embora com desafios em termos de controle de qualidade [Ribeiro et al. 2011]. A *Similarity Mean Opinion Score* (SMOS)³, por sua vez, foca na avaliação da semelhança entre áudios sintetizados e de referência, sendo útil para medir quão próximo um áudio gerado está de uma voz original em termos de características acústicas e vocais [Ren et al. 2021]. Já a *Comparative Mean Opinion Score* (CMOS) avalia a qualidade relativa entre duas versões de áudio sintetizado, pedindo aos avaliadores que comparem diretamente os áudios e apontem qual deles possui melhor qualidade, utilizando uma escala de -3 a +3 [Ren et al. 2022]. Cada uma dessas variantes da MOS foca em diferentes aspectos da qualidade de áudio, utilizadas de acordo com o que se deseja avaliar. Considerando que estudos têm utilizado a MOS como uma medida de naturalidade da fala em tarefas de síntese (cf. [Sellam et al. 2023], [Choi et al. 2022]), a descrição da característica observada pelo avaliador foi adaptada para a avaliação de naturalidade (veja a coluna 4 da Tabela 1).

²Acesso em <https://github.com/GustavoEvangelistaAraujo/EyetrackingMOS-STIL>

³Também abreviado por SimMOS na literatura.

3. EyetrackingMOS

O rastreamento ocular é amplamente reconhecido como uma das técnicas mais precisas para a avaliação *online* do processamento linguístico [Mitchell 2004, Kaiser 2013]. Os variados movimentos dos olhos durante o processamento de informações podem ser utilizados para inferir como essas informações são processadas, seja durante a leitura de texto (estímulo de leitura) ou ao observar uma imagem (estímulo visual). O resultado é obtido a partir da porcentagem de tempo em que o avaliador olhou para o lado direito e esquerdo, os quais mostram figuras relacionadas aos conceitos que se deseja medir. Assim, registramos a porcentagem de tempo que o participante permanece com o olhar sobre a figura que representa a fala natural. Esta medida pode ser avaliada em um intervalo de 0% a 100% e mapeada para a escala MOS como apresentado na Tabela 1. Assim como no MOS, ao final é calculada a média das notas de todos os avaliadores.

Tabela 1. Mapeamento entre pontuações do EyetrackingMOS e MOS

Tempo de fixação (%)	Avaliação MOS	Qualidade	Naturalidade
81 a 100	5	Excelente	Extremamente natural
61 a 80	4	Boa	Muito natural
41 a 60	3	Razoável	Razoavelmente natural
21 a 40	2	Pobre	Pouco natural
0 a 20	1	Ruim	Nada natural

4. Materiais e métodos

4.1. Descrição do conjunto de dados

Há uma carência de conjuntos de dados de áudio com variantes linguísticas regionais. Corpus como BRACCENT, utilizado em [Batista 2019], [Ling et al. 2018] e [Ynoguti 1999], não apresentam volume satisfatório de dados, assim como não tratam da fala espontânea. Portanto, foi escolhido para este estudo um recorte de áudios de um grande dataset do Museu da Pessoa⁴, um museu virtual e colaborativo de histórias de vida, que são do tipo entrevistas biográficas, compilado pelo projeto Tarsila⁵. Detalhes do recorte preliminar do corpus (MuPe-v1) estão disponíveis no repositório².

4.2. Modelo de síntese de fala

O modelo SYNTACC [Nguyen et al. 2023] é uma arquitetura para síntese de fala com múltiplos sotaques baseada no YourTTS [Casanova et al. 2022]. Similarmente ao antecessor, utiliza uma arquitetura de codificação-decodificação baseada em Transformer, onde o codificador recebe a sequência de texto como entrada e gera uma representação intermediária, que é posteriormente processada pelo decodificador para gerar o espectrograma mel, uma representação em espectro da frequência ao longo do tempo que é reconstruída em áudio por um *vocoder*.

Esse modelo implementa as seguintes mudanças: na entrada, a arquitetura concatena 4 *embeddings* de idiomas treináveis em cada caractere de entrada, uma técnica de fatorização de pesos (*weight factorization*), o que permite um treinamento *multi-accent*.

⁴<https://museudapessoa.org/>

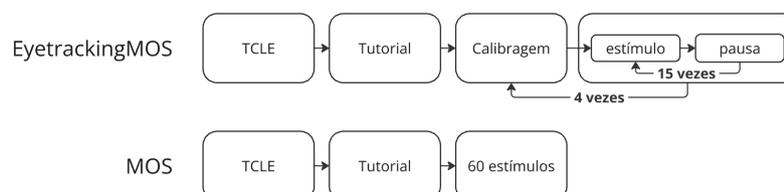
⁵<https://sites.google.com/view/tarsila-c4ai/home>

Esta abordagem divide os pesos do modelo em componentes compartilhados e específicos para cada variante linguística, otimizando o treinamento em cenários de poucos recursos. Isso possibilita que a síntese de fala seja adaptável para o contexto do português brasileiro, sendo possível obter um controle explícito de sotaques pelo congelamento parcial de pesos atribuídos a ele e portanto permite que a fala seja sintetizada de forma mais específica para cada variante. Detalhes da arquitetura, configurações do modelo e etapa de treinamento também foram disponibilizados no repositório².

5. Experimentos

A Figura 1 apresenta o fluxo de interação do usuário neste experimento. Para tanto, foi utilizada a plataforma Gorilla⁶, uma plataforma paga, com o objetivo de construção e coleta de tarefas de anotação. A sequência de interfaces e o conjunto de dados dos estímulos também são apresentados no repositório².

Figura 1. Fluxograma da interação do usuário em cada experimento



No experimento elaborado neste trabalho, o processo se inicia com a aceitação do Termo de Consentimento Livre e Esclarecido (TCLE). Em seguida, o participante é conduzido a um tutorial, que tem o objetivo de ambientá-lo com o experimento subsequente. Para a captura do vídeo, são utilizadas as câmeras padrões dos dispositivos pessoais (apenas computador e notebook) dos usuários, caso as configurações de iluminação e qualidade de imagem não sejam suficientemente boas para permitir que o participante complete a calibragem sem erros, o participante é impedido de continuar. Em conjunto com uma calibragem recorrente, é possível inferir que a qualidade de rastreamento se mantenha desde o início até o final do experimento. O Gorilla utiliza a biblioteca Webgazer⁷ para rastreamento ocular. No caso do EyetrackingMOS, após o tutorial, o usuário passa pela etapa de calibragem, que é dividida em duas partes. Primeiro, é necessário posicionar corretamente o rosto em relação à câmera. Em seguida, o participante deve fixar o olhar em uma sequência de pontos que aparecem aleatoriamente nas extremidades da área útil da tela. São apresentados 10 pontos no total, no qual os 5 primeiros pontos tornam-se uma referência de rastreamento, e os 5 seguintes são repetidos como validação dos anteriores. Caso haja uma discrepância significativa entre a referência e a validação em um dos pontos (considerado como a tolerância do teste), a calibragem é considerada falha, e o usuário precisa repetir o processo. Após uma calibragem bem-sucedida, o participante prossegue e visualiza uma tela com duas imagens vetoriais ilustrativas (um robô e uma figura humana, que trocam de posição de forma aleatória a cada estímulo) enquanto ouve o áudio. A pausa é uma tela subsequente ao final do áudio com apenas um sinal de “+” por 3 segundos, feita para poder reposicionar o olhar do usuário no meio da tela.

⁶<https://gorilla.sc/>

⁷<https://webgazer.cs.brown.edu/>

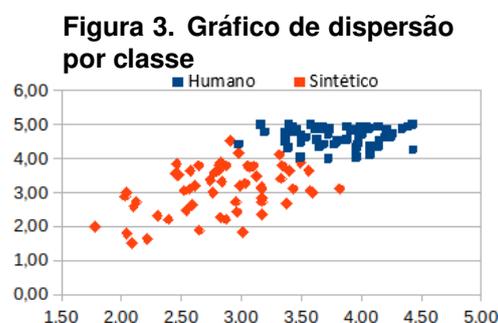
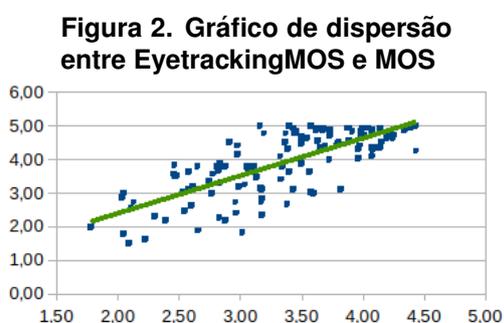
Este ciclo de estímulo e pausa é repetido 15 vezes, e então é feita uma nova calibragem para garantir a qualidade de rastreamento do globo ocular, sendo realizado quatro ciclos completos, que totalizam 60 estímulos.

Por outro lado, no experimento de MOS, após o TCLE e o tutorial, o participante é exposto a 60 estímulos de áudio. Durante o tutorial, são apresentados três exemplos de áudios sintetizados, correspondentes às pontuações 1, 3 e 5, para ajudar o participante a alinhar suas expectativas. O participante pode ouvir cada estímulo mais de uma vez antes de decidir sua pontuação, utilizando a Tabela 1 como referência para todas as 60 amostras de áudio, conforme descrito na literatura de avaliação de modelos de síntese.

A divisão das listas de áudios para avaliação foi realizada considerando dois tipos de áudios: sintetizados e naturais. Esses áudios foram organizados em duas listas: Lista A e Lista B, que foram atribuídas aos participantes de forma equilibrada. 30 áudios naturais foram colocados na Lista A, enquanto seus correspondentes sintetizados foram alocados na lista B. Da mesma forma, 30 áudios sintetizados foram incluídos na Lista A, com os seus correspondentes naturais na lista B. Quanto aos participantes, conforme [Loizou 2011], a proporção de avaliações subjetivas deve ser de 10 especialistas para 20 não especialistas. Foram escolhidos 76 anotadores dentre 28 especialistas e 48 não especialistas, distribuídos entre 4 grupos de 19 participantes. Ambos experimentos foram elaborados desta mesma forma, o que assegurou uma diversidade de perspectivas nas avaliações, permitindo uma análise comparativa abrangente entre as opiniões de especialistas e não especialistas sobre os áudios apresentados.

6. Resultados preliminares

A Figura 2 ilustra a relação entre os valores mensurados pelo EyetrackingMOS, convertidos para a escala MOS, e os valores mensurados diretamente pelo MOS. Cada ponto azul representa um par de medidas, com o eixo horizontal correspondendo aos valores do EyetrackingMOS convertidos para a escala MOS e o eixo vertical representando os valores obtidos diretamente pelo MOS. A linha verde traçada no gráfico indica a linha de tendência linear, mostrando a direção geral da correlação entre as duas variáveis. A Figura 3 ilustra a dispersão das pontuações obtidas tanto pelo MOS quanto pelo rastreamento ocular (EyetrackingMOS) para áudios reais e sintetizados. Em ambos os testes, os participantes conseguiram separar razoavelmente bem os áudios reais dos sintetizados.



No gráfico de dispersão por MOS (Figura 4), observa-se uma distinção clara entre os áudios reais, que tendem a receber pontuações mais altas, e os sintetizados, que

se concentram nas faixas intermediárias e baixas. No entanto, no gráfico de dispersão por rastreamento ocular (Figura 5), a separação entre áudios reais e sintetizados é menos evidente. Essa maior dispersão nos resultados do rastreamento ocular é esperada, já que nesse método os estímulos são percebidos apenas uma vez, enquanto nos métodos *offline* como o MOS, o anotador pode ouvir o estímulo repetidas vezes antes de tomar sua decisão, resultando em uma separação mais clara entre os tipos de áudio. Assim, o rastreamento ocular oferece uma avaliação mais detalhada, capturando variações mais sutis na percepção da qualidade dos áudios.

Figura 4. Gráfico de dispersão por MOS

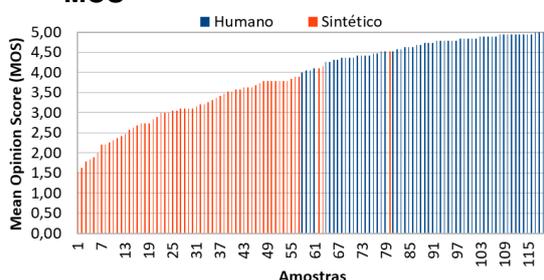
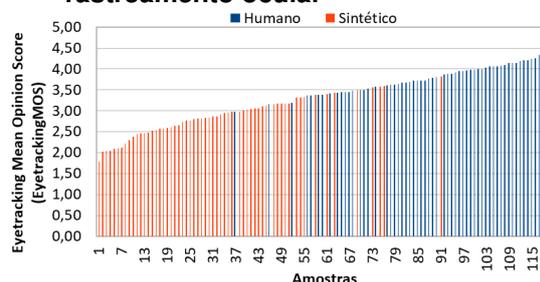


Figura 5. Gráfico de dispersão por rastreamento ocular



Os resultados apresentados na Tabela 2 indicam uma correlação razoável entre o EyetrackingMOS e o MOS, com uma métrica R^2 de 56%, sugerindo que o EyetrackingMOS explica 56% da variância observada no MOS. O desvio padrão do erro entre as duas métricas é de 0,72 unidades, mostrando que, em geral, elas tendem a ser próximas, com uma diferença média de menos de uma unidade. Além disso, o MOS tende a classificar um número maior de áudios com a nota máxima ou valores próximos, enquanto o EyetrackingMOS oferece uma análise mais detalhada, por sua escala ser de 0 a 100, o que é observado em áudios de alta qualidade. Essa dispersão indica que, embora exista uma correlação razoável entre as duas métricas, conforme evidenciado pela inclinação positiva da linha de tendência, as medidas não são perfeitamente alinhadas, refletindo diferenças na maneira como cada método capta e avalia a qualidade dos áudios.

Tabela 2. Medidas de performance estatística

Medida	Valor	Interpretação geral
Pearson	0.744	Correlação moderada
Mean Squared Error (MSE)	0.710	Erro médio baixo
Rooted Mean Squared Error (RMSE)	0.844	Erro médio baixo
R^2	0.553	Explica 55% da variância
Spearman	0.714	Correlação moderada

Também foi realizada uma análise da concordância entre os avaliadores dentro de seus respectivos grupos, utilizando o coeficiente de Kendall's W para avaliar a consistência das respostas (Tabela 3). Em resumo, o grupo EyetrackingMOS apresentou maior consistência nas avaliações, com alta concordância na maioria dos estímulos, enquanto o grupo MOS demonstrou uma maior variabilidade, com concordância que variou de alta até nenhuma, indicando possíveis desafios na avaliação uniforme dos

estímulos por este grupo. Com relação ao tempo, EyetrackingMOS e MOS tomaram em média 12:07min e 12:30min dos participantes, respectivamente. As medianas foram de 11:38min e 10:41min, respectivamente. Nota-se que o teste MOS tende a ser em torno de 1 minuto mais rápido que o EyetrackingMOS que pode ser justificada pelo tempo das 4 calibrações do rastreamento ocular.

Tabela 3. Medidas de concordância para cada grupo de experimentos

Grupo	Intervalo de Kendall's W	Interpretação geral
EyetrackingMOS	0.6719 a 0.9579	Alta concordância geral, algumas variações
MOS	0.0000 a 0.9474	Grande variação, alta concordância a nenhuma concordância

7. Conclusão e trabalhos futuros

Conforme os resultados preliminares, o EyetrackingMOS e MOS têm uma correlação razoável. Paralelamente, a utilização de uma medida de avaliação subjetiva com rastreamento ocular oferece vantagens significativas, uma vez que permite capturar reações genuínas e síncronas aos estímulos apresentados. Além disso, o controle mais rigoroso sobre a quantidade de estímulos recebidos por cada participante pode reduzir a variação na concordância e aumentar a quantidade de estímulos por sessão. A reação mecânica ocular também pode reduzir variações na concordância, causadas pelas diferentes interpretações das descrições de pontuação de métricas subjetivas. A escala de 0 a 100 para cada indivíduo oferece uma avaliação mais detalhada e precisa, permitindo uma maior granularidade na análise das respostas, ao contrário das escalas limitadas a poucos pontos. Embora a produção dessa medida seja mais complexa e demorada, o benefício de obter uma análise mais transparente das reações dos participantes justifica seu uso como complemento do MOS tradicional.

Como trabalhos futuros, pretende-se experimentar diferentes tecnologias/plataformas de captação ocular para comparar a precisão da captação. Também é importante obter dados estatísticos com uma distinção das pontuações fornecidas entre os grupos de especialistas e não especialistas. Além disso, a seleção de variáveis deve ser refinada, como, por exemplo, calcular a fixação no espaço intermediário entre as imagens, o que pode oferecer uma compreensão mais detalhada das reações dos participantes. Por fim, explorar maneiras de realizar esses testes gratuitamente, seja por meio de parcerias, uso de plataformas de *crowdsourcing* ou outras abordagens que reduzam os custos e ampliem o acesso aos participantes.

8. Agradecimentos

Este trabalho foi realizado no Centro de Inteligência Artificial (C4AI-USP), com o apoio da Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP, bolsa #2019/07665-4) e da IBM Corporation. O projeto também foi apoiado pelo Ministério da Ciência, Tecnologia e Inovações, com recursos da Lei nº 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. Agradecimentos também são dirigidos ao Programa de Excelência Acadêmica (PROEX) da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES), nº 88887.841258/2023-00.

Referências

- ALMEIDA, R. A. S. d., OLIVEIRA JR., M., and COZIJN, R. (2021). *Paradigma do Mundo Visual: Método de Rastreamento Ocular*, chapter 5. Blucher Open Access.
- Batista, N. A. R. (2019). Estudo sobre identificação automática de sotaques regionais brasileiros baseada em modelagens estatísticas e técnicas de aprendizado de máquina. Master's thesis, Unicamp.
- Cagliari, L. C. (1992). Prosódia: algumas funções dos supra-segmentos. *Cadernos de estudos linguísticos*, 23:137–151.
- Casanova, E., Shulby, C., Gölge, E., Müller, N. M., de Oliveira, F. S., Junior, A. C., da Silva Soares, A., Aluisio, S. M., and Ponti, M. A. (2021). Sc-glowtts: an efficient zero-shot multi-speaker text-to-speech model.
- Casanova, E., Weber, J., Shulby, C. D., Junior, A. C., Gölge, E., and Ponti, M. A. (2022). Yourtts: Towards zero-shot multi-speaker tts and zero-shot voice conversion for everyone. In *International Conference on Machine Learning*, pages 2709–2720. PMLR.
- Caseli, H. M. and Nunes, M. G. V., editors (2024). *Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português*. BPLN, 2 edition.
- Choi, Y., Jung, Y., Suh, Y., and Kim, H. (2022). Learning to maximize speech quality directly using mos prediction for neural text-to-speech. *IEEE Access*, 10:52621–52629.
- Cooper, E., Huang, W.-C., Tsao, Y., Wang, H.-M., Toda, T., and Yamagishi, J. (2024). A review on subjective and objective evaluation of synthetic speech. *Acoustical Science and Technology*, 45(4):161–183.
- Hoogeboom, E., Van Den Berg, R., and Welling, M. (2019). Emerging convolutions for generative normalizing flows. In *International conference on machine learning*, pages 2771–2780. PMLR.
- ITU - R (2017). ITU-T Rec. P.10/G.100 (11/2017): Vocabulary for performance, quality of service and quality of experience. Recommendation P.10/G.100, International Telecommunication Union. <https://www.itu.int/rec/T-REC-P.10-201711-I/en>.
- ITU - T (1996). Methods for subjective determination of transmission quality. Recommendation P.800, International Telecommunication Union.
- Jia, Y., Zhang, Y., Weiss, R. J., Wang, Q., Shen, J., Ren, F., Chen, Z., Nguyen, P., Pang, R., Moreno, I. L., and Wu, Y. (2019). Transfer learning from speaker verification to multispeaker text-to-speech synthesis.
- Ju, Z., Wang, Y., Shen, K., Tan, X., Xin, D., Yang, D., Liu, Y., Leng, Y., Song, K., Tang, S., Wu, Z., Qin, T., Li, X.-Y., Ye, W., Zhang, S., Bian, J., He, L., Li, J., and Zhao, S. (2024). NaturalSpeech 3: Zero-shot speech synthesis with factorized codec and diffusion models.
- Kaiser, E. (2013). Experimental paradigms in psycholinguistics. In Podesva, R. J. and Sharma, D., editors, *Research Methods in Linguistics*, pages 135–168. Cambridge University Press, Cambridge.

- Kingma, D. P., Salimans, T., Jozefowicz, R., Chen, X., Sutskever, I., and Welling, M. (2016). Improved variational inference with inverse autoregressive flow. *Advances in neural information processing systems*, 29.
- Le Maguer, S., King, S., and Harte, N. (2024). The limits of the mean opinion score for speech synthesis evaluation. *Computer Speech Language*, 84:101577.
- Ling, L., Fernandes Tavares, T., Barbosa, P., and Batista, N. (2018). Detecção automática de sotaques regionais brasileiros: A importância da validação cross-datasets.
- Loizou, P. C. (2011). *Speech Quality Assessment*, pages 623–654. Springer Berlin Heidelberg, Berlin, Heidelberg.
- Mitchell, D. C. (2004). On-line methods in language processing: introduction and historical review. In Carreiras, M. and Clifton Jr., C., editors, *The On-line Study of Sentence Comprehension: Eyetracking, ERP and Beyond*, pages 15–32. Psychology Press.
- Mota, J. A., Ribeiro, S. S. C., and de Oliveira, J. M. (2023). *Atlas Linguístico Do Brasil: Comentários às Cartas Linguísticas 1-V. 3*. Ed. Universidade Estadual de Londrina.
- Nguyen, T.-N., Pham, N.-Q., and Waibel, A. (2023). Syntacc: Synthesizing multi-accent speech by weight factorization. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.
- Ren, Y., Hu, C., Tao, X., Zhao, Z., Zhang, X., Li, Q., Lei, L., Zhou, S., Liu, J., and Liu, S. (2021). Fastspeech 2: Fast and high-quality end-to-end text to speech. In *International Conference on Learning Representations*.
- Ren, Y., Zhao, Z., Tan, X., Yi, J., Cheng, Y.-L., Yang, J., Qin, T., and Liu, T.-Y. (2022). Naturalspeech: End-to-end text to speech synthesis with human-level quality. In *Advances in Neural Information Processing Systems*.
- Ribeiro, F., Florêncio, D., Zhang, C., and Seltzer, M. (2011). Crowdmos: An approach for crowdsourcing mean opinion score studies. In *2011 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 2416–2419. IEEE.
- Sellam, T., Bapna, A., Camp, J., Mackinnon, D., Parikh, A. P., and Riesa, J. (2023). Squid: Measuring speech naturalness in many languages. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.
- Shen, K., Ju, Z., Tan, X., Liu, Y., Leng, Y., He, L., Qin, T., Zhao, S., and Bian, J. (2023). Naturalspeech 2: Latent diffusion models are natural and zero-shot speech and singing synthesizers.
- Tan, X., Chen, J., Liu, H., Cong, J., Zhang, C., Liu, Y., Wang, X., Leng, Y., Yi, Y., He, L., Soong, F., Qin, T., Zhao, S., and Liu, T.-Y. (2022). Naturalspeech: End-to-end text to speech synthesis with human-level quality.
- Ynoguti, C. A. (1999). *Reconhecimento de Fala Contínua Utilizando Modelos Ocultos de Markov*. PhD thesis, Unicamp.



Quati: A Brazilian Portuguese Information Retrieval Dataset from Native Speakers

Mirelle Bueno^{*1}, E. Seiti de Oliveira^{*1}, Rodrigo Nogueira^{1,2}, Roberto Lotufo^{1,3}, Jayr Pereira⁴

¹Departamento de Engenharia de Computação e Automação (DCA)
Universidade Estadual de Campinas – UNICAMP
13083-872 – Campinas – São Paulo, Brasil

²Maritaca AI

³NeuralMind.ai

⁴Universidade Federal do Cariri
Juazeiro do Norte-CE, Brasil.

m174909@dac.unicamp.br, eduseiti@dca.fee.unicamp.br

{rfn,lotufo}@unicamp.br, jayr.pereira@ufca.edu.br

Abstract. We present *Quati*,¹ a dataset specifically designed for evaluating Information Retrieval (IR) systems for the Brazilian Portuguese language. It comprises a collection of queries formulated by native speakers and a curated set of documents sourced from a selection of frequently accessed Brazilian Portuguese websites, which ensures a representative and relevant corpus. To label the query–document pairs, we use a state-of-the-art LLM, which shows inter-annotator agreement levels comparable to human performance in our assessments. Our annotation methodology is described, enabling the cost-effective creation of similar datasets for other languages, with an arbitrary number of labeled documents per query. As a baseline, we evaluate a diverse range of open-source and commercial retrievers. *Quati* is publicly available at <https://huggingface.co/datasets/unicamp-dl/quati>, and all scripts at <https://github.com/unicamp-dl/quati>.

1. Introduction

The development of Information Retrieval (IR) systems depends on high-quality evaluation datasets, which should contain queries and documents ideally in the same target language of those systems, in order to capture specific information needs and social-cultural aspects. That, contrasts with translated datasets, which potentially represent the information needs and knowledge of a different culture or society. Hence, translated datasets may not effectively measure a retrieval system’s ability in real-world scenarios involving native users.

^{*}Equal contribution.

¹We named our dataset after this South American mammal, whose foraging behavior represents the resolute search for resources.

Despite being one of the most widely spoken languages in the world, there is a scarcity of IR datasets in Portuguese. Existing datasets such as REGIS [Lima de Oliveira et al. 2021] and RCV1 [Lewis et al. 2004]², though valuable, fall short due to their limited size and specialized domains (geoscience and news). While translated datasets such as mMARCO [Bonifacio et al. 2021] and mRobust04 [Jeronymo et al. 2022] have helped to alleviate this issue, the use of automatic translations often represents the loss of socio-cultural characteristics of the target languages, and the evaluations may become biased by the source language.

To address those issues, we created Quati, a Brazilian Portuguese evaluation dataset, comprising human-written queries and a high-quality native corpus. Quati is created using a semi-automated pipeline, aiming to reduce the labeling cost barrier. We use a Large Language Model (LLM) to judge a passage’s relevance for a given query, publishing a cost-effective pipeline to create an IR evaluation dataset with an arbitrary number of annotated passages per query.³ In this context, our work aims to answer the following research question: Can LLMs be used to compose a semi-automated pipeline for annotating query–passages relevance for Brazilian Portuguese IR systems?

To evaluate the quality of the LLM annotations, we compare them with human annotations on a sample of query–passage pairs and confirmed a Cohen’s Kappa coefficient of 0.31. While this figure is below the 0.41 seen in human-human annotation agreement, it is consistent with the findings reported in the literature [Faggioli et al. 2023, Thomas et al. 2023, Farzi and Dietz 2024] and it will likely increase as LLMs improve in quality. The usage of a modular semi-automated pipeline, allows the dataset construction method to be replicated to create high-quality IR datasets for other languages.

2. Related Work

Evaluation datasets are an important variable in the IR context as they expose the limitations of search systems and guide their development. However, most of the available datasets are in English, as is the case with MS MARCO [Bajaj et al. 2016]. Works such as MIRACL [Zhang et al. 2023], mMARCO [Bonifacio et al. 2021], mRobust [Jeronymo et al. 2022], Mr.Tydi [Clark et al. 2020], TREC CLIR [Schäuble and Sheridan 1998], CLEF [Peters and Braschler 2002], NT-CIR [Sakai et al. 2021] and HC4 [Lawrie et al. 2022] are efforts to develop datasets for other languages, but most are based on language translation to adapt English to the target languages, or do not include Portuguese. Ongoing efforts [Lima de Oliveira et al. 2021, Vitória et al. 2024] are starting to change that scenario creating IR datasets for Brazilian Portuguese, but so far focusing on specific domains.

The creation of datasets for IR is a resource-intensive task, particularly in the process of judging the relevance of documents. Recent endeavors have witnessed a shift towards leveraging LLMs to assess query–passage relevance [Zendel et al. 2024]. Faggioli et al. [Faggioli et al. 2023] further underscored the potential of employing LLMs for automating the judgment of document relevance, thereby opening up promising avenues

²<https://trec.nist.gov/data/reuters/reuters.html>

³The total cost for this dataset was US\$140.19 (0.03 per query–passage) for an average of 97.78 annotated passages per query.

for exploration in this domain. Complementary evaluations conducted by Thomas et al. [Thomas et al. 2023] demonstrated a significant correlation between human judgments and those made by the GPT-3.5-turbo model.

3. Methodology

We used a semi-automatic method to create Quati, as depicted in Figure 1. The required inputs are: 1) A large corpus, originally written in the target language, from which we extract the passages to compose our IR dataset; 2) A set of test queries, manually created to represent the information needs of native speakers. In the following sections, we detail the steps of the pipeline.

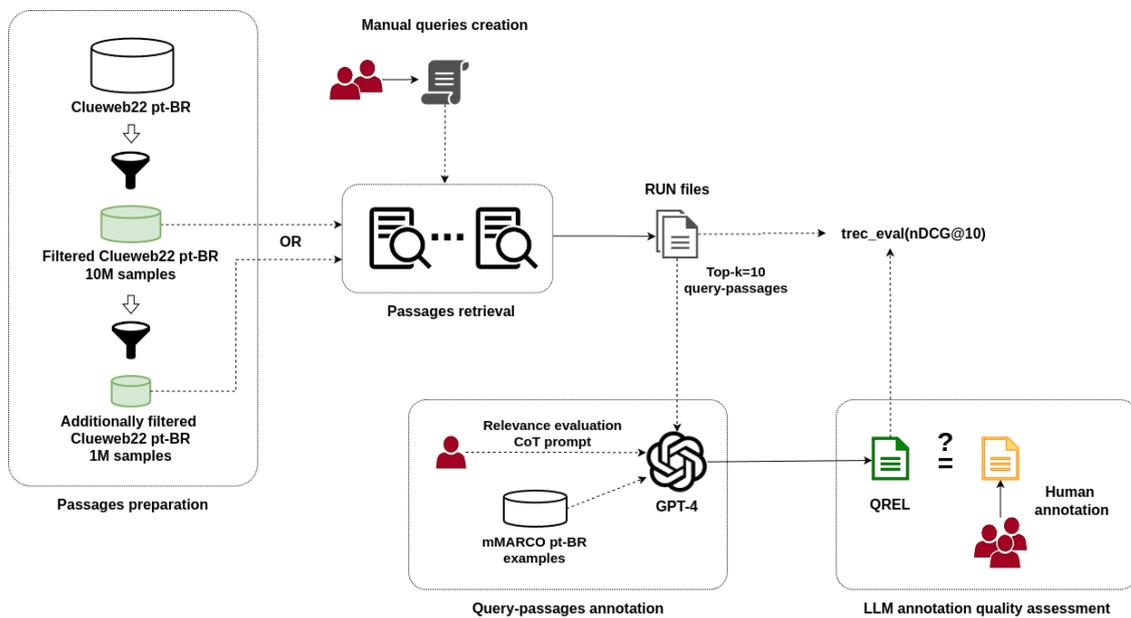


Figure 1. Proposed IR dataset creation methodology.

3.1. Passages preparation

The passages preparation step is composed by the following substeps:

Data collection: We used the Portuguese subset of ClueWeb22 [Overwijk et al. 2022] category B, which includes 4.1 million web pages more likely to be visited according to Bing search algorithms during the first half of 2022 [Overwijk et al. 2022].

URL filtering: We excluded any documents from our dataset whose URLs’ domain ended with “.pt”, which refer to Portuguese from Portugal, as the language style in those documents might differ significantly from that used in Brazilian Portuguese web pages. Additionally, we used FastText [Joulin et al. 2016b, Joulin et al. 2016a] as an additional language verification method to ensure that only Portuguese documents were included in our corpus.

Document segmentation into passages: Following language verification, we segmented the documents into approximately 1,000-character segments and assessed the percentage of line breaks ($\backslash n$) occurrences within each segment, removing those with more than 20%.

This criterion was used to increase the probability of retaining segments predominantly composed of natural language text.

Separation into large and small versions: With the process described in the previous steps, we collected a total of 20 million segments. From this set, we randomly selected 10 million segments (hereinafter referred to as 10M corpus) to be the passages in our corpus, creating a large, but still manageable, dataset of more than 11 GB of size. A second dataset was built from the first one applying additional filtering rules taken from the MassiveWeb Corpus [Rae et al. 2021] — e.g. removing passages with more than 10% of symbols, or with mean word length outside the 3 to 10 interval — and sampling only 1 million segments (hereinafter referred to as 1M corpus) from the resulting 7M filtered documents — the goal was to create a smaller and higher-quality dataset that would facilitate experimentation with embedding models, as encoding the original 10 million segments can be computationally expensive.

3.2. Manual queries creation

We employed human-created queries for the evaluation dataset, aiming high-quality questions to capture common information needs from a diverse corpus, created by native speakers of the target language. We created a total of 200 test queries.

3.3. Passages retrieval

The next step is the passage retrieval to build a list of query–passages to annotate. As it would be prohibitive to have the relevance scores each query for the entire corpus, we select to annotate the top-k passages returned by multiple IR systems. It is assumed that the diversity of their results will enable the collection of a variety of passages, creating a robust evaluation dataset.

We selected a mix of strong and weak IR systems, to include a variety of passages: **BM25**: a strong baseline for retrieval; **BM25 + mT5-XL**: two-stage pipeline with BM25 followed by mT5-XL (3.7 billion parameters) [Xue et al. 2020]; **BM25 + E5-large**: two-stage pipeline with BM25 and E5-large [Wang et al. 2022]⁴; **E5-large** and **E5-base**: E5 variants as dense retrievers, using FAISS [Johnson et al. 2019] with inner product for search; **ColBERT-X** [Nair et al. 2022]: a multilingual ColBERT-v1 fine-tuned in Brazilian Portuguese subset of mMARCO; **SPLADE v2**: a learned sparse retriever [Formal et al. 2021] fine-tuned on Brazilian Portuguese subset of mMARCO; **SPLADE v2 + mT5-XL**: two-stage pipeline using SPLADE v2 followed by mT5.

We also use Reciprocal Ranking Fusion (RRF) [Cormack et al. 2009] to increase the retrieved documents diversity, using the following combinations: **E5-large + ColBERT-X**; **E5-large + SPLADE v2**; and **E5-large + BM25 + mT5-XL**.

We also include commercial embedding models: **text-embedding-ada-002**⁵, **text-embedding-3-small**⁶ and as it employs the Matryoshka Representation Learning technique [Kusupati et al. 2022], we performed the retrieval using only the first half dimensions (identified as text-embedding-3-small half). FAISS [Johnson et al. 2019] using inner product was applied for dense vectors search for all of them.

⁴<https://huggingface.co/intfloat/multilingual-e5-large>

⁵<https://openai.com/blog/new-and-improved-embedding-model>

⁶<https://openai.com/blog/new-embedding-models-and-api-updates>

To evaluate the diversity of retrieved passages, we counted the query–passage combinations exclusively returned by each IR system, which should be a number from 0 to 500, 0 meaning the query–passages returned by a particular IR system were also returned by another IR system. Although we look for diversity, there should be a balance: we could have reached 5,000 different query–passage combinations (10 IR systems, 50 queries, 10 passage/query) if all systems returned exclusive passages, but that would indicate no agreement on the most relevant passages per query.

3.4. Query–passages annotation

The final step of the query annotation is to use an LLM to label the retrieved passages’ relevance for each query. We selected the top-k=10 passages for a sample of 50 queries using all the retrieval systems considered on both the 10M and 1M corpora and sent them for LLM evaluation. We applied a few-shot Chain-of-Thought (CoT) prompt [Wei et al. 2022], and we adopted the TREC 2021 Deep Learning track 4-score relevance annotation scale [Craswell et al. 2021]: (1) **Irrelevant**: the passage is outside the scope of the question; (2) **Relevant**: the passage pertains to the question’s topic but does not provide a direct answer; (3) **Highly relevant**: the passage answers the question, but lacks in clarity or has unrelated information. (4) **Perfectly relevant**: the passage answers the question with clarity and precision.

We selected OpenAI GPT-4 model⁷ as the annotator. Due to cost limitations, we used a 50-sample from the 200 queries. We asked the LLM to label only the top-10 retrieved passages of each IR system for each query. We used a CoT prompt with two in-context examples selected from the mMARCO pt-BR dataset [Bonifacio et al. 2021]. The prompt written in Brazilian Portuguese includes the task explanation and the CoT examples to produce the 4-score passage relevance value for a given query. The final evaluation was requested in JSON format to simplify the LLM response parsing process. The prompt was built and refined using a limited set of questions sampled from the same mMARCO pt-BR dataset. The final prompt version can be found online.⁸

4. Experiments

4.1. LLM annotation quality assessment

We assess the quality of our LLM-based annotator by comparing its query–passage relevance scores with those provided by human annotators. This process was conducted on a 24-sample of the 50 annotated queries. Using the Doccano [Nakayama et al. 2018] system, three researchers annotated the top-10 passages returned by the BM25 + mT5 IR system applying the same TREC-DL 2021 4-score grading system. The agreement among the query–passage relevance annotations generated by the LLM and humans was measured using Cohen Kappa, Pearson, and Spearman correlation coefficients.

4.2. Retrieval systems evaluation

We used the LLM annotated query–passages to evaluate the IR systems effectiveness in the 10M and 1M Quati datasets. As we already have the IR runs for the passages retrieval by all the systems (see Section 3.3), we simply compute the nDCG@10 metric over those

⁷We used gpt-4-1106-preview, available at the OpenAI API.

⁸<https://github.com/unicamp-dl/quati/blob/main/prompt.md>

Table 1. The single-system query-passages column indicates the ones returned only by that system, either for the 10M or the 1M sets; the percentage refers to 500 query-passages. For the single system total, the percentage refers to the union of evaluated passages. “Others” are results with data preparation issues, but valid annotations.

Retrieval System	Single-system query-passages	
	10M dataset	1M dataset
E5-base	262 (52.4%)	
BM25	248 (49.6%)	253 (50.6%)
SPLADE v2 pt-BR	151 (30.2%)	
E5-large	122 (24.4%)	
ColBERT-X mMARCO pt-BR	115 (23.0%)	195 (39.0%)
BM25 + E5-large	115 (23.0%)	120 (24.0%)
SPLADE v2 pt-BR + mT5-XL	86 (17.2%)	
BM25 + mT5-XL	60 (12.0%)	93 (18.6%)
E5-large + ColBERT-X mMARCO pt-BR RRF	32 (6.4%)	
E5-large + SPLADE v2 pt-BR RRF	29 (5.8%)	
text-embedding-ada-002		137 (27.4%)
text-embedding-3-large		121 (24.2%)
text-embedding-3-small half		45 (9.0%)
text-embedding-3-small		31 (6.2%)
Others	814 (54.27%)	
Single system query-passages total	3029 (61.96%)	
Union of all systems query-passages	4889	

results. Besides establishing a baseline for a variety of IR systems, this experiments also indirectly assess the overall quality of Quati validation dataset: by verifying different effectiveness for already published IR systems, we validate Quati potential to indeed assess such systems.

5. Results and Discussion

5.1. Annotated passages variability

Table 1 indicates a range from 29 to 262 query–passage combinations exclusively returned by a single IR system. On average, each system returned 28.85% of new passages, and from the total 4,889 evaluated query–passages, 61.96% (3029) were returned by a single system, suggesting our pool of systems is diverse. As shown in Table 2, the IR systems were able to retrieve a diversity set of query–passages, including “perfectly relevant” (score=3) ones; also, the diversity increased for less relevant passages, indicating the systems agreed more as the passage relevance increased.

5.2. LLM annotations quality is aligned with crowd workers

Table 3 shows the Cohen’s Kappa and the Spearman’s Rho correlation coefficients for the human and LLM annotations, computed for the 240 query–passage combinations. The average Cohen’s Kappa of 0.31 is aligned with the literature. For example,

Table 2. Query–passage relevance score counts. The systems agreed more, returning the same passages per query, as the relevance score increases. “Relevant” includes passages from scores 1 to 3.

Score	All query–passages	Single-system query–passages	%
0	2489	1839	73.89
1	985	586	59.49
2	759	375	49.41
3	656	229	34.91
Relevant	2400	1190	49.58
Total	4889	3029	61.96

Table 3. Cohen’s Kappa and Spearman’s Rho correlations among Human Annotators (HA) and the GPT-4, for the query–passage 4-score evaluations. For each annotator, 4th row holds the average of the correlation against the others. We then compute the mean of that value only for the Human Annotators (“Mean HA” row), to characterize their overall correlation.

	Cohen’s Kappa				Spearman’s Rho			
	HA ₁	HA ₂	HA ₃	GPT-4	HA ₁	HA ₂	HA ₃	GPT-4
HA ₁	—	0.4369	0.4294	0.3234	—	0.6931	0.6924	0.6073
HA ₂	0.4369	—	0.4105	0.2593	0.6931	—	0.6985	0.6174
HA ₃	0.4294	0.4105	—	0.3498	0.6924	0.6985	—	0.6296
Mean	0.4331	0.4237	0.4199	0.3108	0.6927	0.6958	0.6954	0.6181
Std	0.0037	0.0132	0.0095	0.0380	0.0004	0.0027	0.0031	0.0091
Mean HA	0.4256±0.0055			—	0.6946±0.0014			—
Diff. Mean HA	0.0076	-	-	-	-	0.0011	0.0008	-
		0.0019	0.0057	0.1096	0.0019			0.0765

[Faggioli et al. 2023] reported 0.26 for GPT-3.5, and [Thomas et al. 2023] reported Cohen’s Kappa ranging from 0.20 to 0.64, depending on the prompt used on GPT-4. Our human annotators’ mean Cohen’s Kappa of 0.4256 falls within crowd workers interval of a 0.24 to 0.52, according to [Damessie et al. 2017].

As query–passage relevance annotation is a subjective task, we argue a non-categorical metric such as the Spearman’s Rho would be more appropriate to measure the annotators’ correlation, as errors by a single score level should be considered “less critical”, or within the subjectivity intrinsic for the task. Although human annotators’ correlation is still above their correlations with the LLM, Spearman metrics are within a higher value, better capturing the current LLM effectiveness on the query–passage relevance evaluation.

5.3. Retrieval systems evaluation results

We evaluated the retrievers effectiveness using the LLM annotated query–passages (qrels); table 4 present the results for both the 10M and the 1M datasets. The ranking

Table 4. The nDCG@10 effectiveness on the 50 test queries. The results follows the IR literature and suggests the dataset can effectively evaluate a range of different IR systems.

Retrieval system	nDCG@10	
	10M dataset	1M dataset
BM25	0.4467	0.3991
E5-large	0.5563	—
SPLADE v2 pt-BR	0.5806	—
E5-large + SPLADE v2 pt-BR RRF	0.6272	—
ColBERT-X mMARCO pt-BR	0.6279	0.4927
BM25 + E5-large	0.6364	0.5423
text-embedding-ada-002	—	0.5630
text-embedding-3-small	—	0.5688
text-embedding-3-large	—	0.6319
E5-large + ColBERT-X mMARCO pt-BR RRF	0.6377	—
SPLADE v2 pt-BR + mT5-XL	0.6966	—
BM25 + mT5-XL	0.7109	0.6593

of retrievers with respect to effectiveness matched our expectations, following the literature. We consider that an additional indication of the overall datasets quality as, despite being created in a semi-automated cost-effective way, they are able to evaluate a diversity of retrievers.

6. Conclusion

This paper introduced the Quati, a dataset for supporting the development of IR systems for Brazilian Portuguese retrieval tasks. Quati is publicly available in two sizes, 10M a 1M passages, with 50-query qrels with respectively an average of 97.78 and 38.66 annotated passages per query. Through comparisons with human annotators we answer our research question, showing that state-of-the-art LLM can be used in a semi-automated and cost-effective way to create IR datasets for a specific target language, in the query–passage annotation role, with equivalent performance of humans: LLM annotations correlate with humans’ in similar way human crowd workers annotations do, for a fraction of the cost.

Acknowledgements

We thank Leodécio Braz da Silva Segundo for the valuable support during the human annotation task. We also thank Leonardo Benardi de Avila and Monique Monteiro for the SPLADE v2 retrievals, using the model they trained for Brazilian Portuguese. This research was partially funded by grant 2022/01640-2 from Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP).

References

Bajaj, P., Campos, D., Craswell, N., Deng, L., Gao, J., Liu, X., Majumder, R., McNamara, A., Mitra, B., Nguyen, T., et al. (2016). Ms marco: A human generated machine reading comprehension dataset. *arXiv preprint arXiv:1611.09268*.

- Bonifacio, L., Jeronymo, V., Abonizio, H. Q., Campiotti, I., Fadaee, M., Lotufo, R., and Nogueira, R. (2021). mmarco: A multilingual version of the ms marco passage ranking dataset. *arXiv preprint arXiv:2108.13897*.
- Clark, J. H., Choi, E., Collins, M., Garrette, D., Kwiatkowski, T., Nikolaev, V., and Palomaki, J. (2020). Tydi qa: A benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics*, 8:454–470.
- Cormack, G. V., Clarke, C. L., and Buettcher, S. (2009). Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 758–759.
- Craswell, N., Mitra, B., Yilmaz, E., Campos, D., Voorhees, E. M., and Soboroff, I. (2021). Trec deep learning track: Reusable test collections in the large data regime. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, pages 2369–2375.
- Damessie, T. T., Nghiem, T. P., Scholer, F., and Culpepper, J. S. (2017). Gauging the quality of relevance assessments using inter-rater agreement. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1089–1092.
- Faggioli, G., Dietz, L., Clarke, C., Demartini, G., Hagen, M., Hauff, C., Kando, N., Kanoulas, E., Potthast, M., Stein, B., et al. (2023). Perspectives on large language models for relevance judgment. *arXiv preprint arXiv:2304.09161*.
- Farzi, N. and Dietz, L. (2024). An exam-based evaluation approach beyond traditional relevance judgments. *arXiv preprint arXiv:2402.00309*.
- Formal, T., Lassance, C., Piwowarski, B., and Clinchant, S. (2021). Splade v2: Sparse lexical and expansion model for information retrieval. *arXiv preprint arXiv:2109.10086*.
- Jeronymo, V., Nascimento, M., Lotufo, R., and Nogueira, R. (2022). mrobust04: A multilingual version of the trec robust 2004 benchmark. *arXiv preprint arXiv:2209.13738*.
- Johnson, J., Douze, M., and Jégou, H. (2019). Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547.
- Joulin, A., Grave, E., Bojanowski, P., Douze, M., Jégou, H., and Mikolov, T. (2016a). Fasttext.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.
- Joulin, A., Grave, E., Bojanowski, P., and Mikolov, T. (2016b). Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.
- Kusupati, A., Bhatt, G., Rege, A., Wallingford, M., Sinha, A., Ramanujan, V., Howard-Snyder, W., Chen, K., Kakade, S., Jain, P., et al. (2022). Matryoshka representation learning. *Advances in Neural Information Processing Systems*, 35:30233–30249.
- Lawrie, D., Mayfield, J., Oard, D. W., and Yang, E. (2022). Hc4: A new suite of test collections for ad hoc clir. In *European Conference on Information Retrieval*, pages 351–366. Springer.

- Lewis, D. D., Yang, Y., Russell-Rose, T., and Li, F. (2004). Rcv1: A new benchmark collection for text categorization research. *Journal of machine learning research*, 5(Apr):361–397.
- Lima de Oliveira, L., Romeu, R. K., and Moreira, V. P. (2021). Regis: A test collection for geoscientific documents in portuguese. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2363–2368.
- Nair, S., Yang, E., Lawrie, D., Duh, K., McNamee, P., Murray, K., Mayfield, J., and Oard, D. W. (2022). Transfer learning approaches for building cross-language dense retrieval models. In *European Conference on Information Retrieval*, pages 382–396. Springer.
- Nakayama, H., Kubo, T., Kamura, J., Taniguchi, Y., and Liang, X. (2018). doccano: Text annotation tool for human. Software available from <https://github.com/doccano/doccano>.
- Overwijk, A., Xiong, C., and Callan, J. (2022). Clueweb22: 10 billion web documents with rich information. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3360–3362.
- Peters, C. and Braschler, M. (2002). The importance of evaluation for cross-language system development: the clef experience. In *LREC*. Citeseer.
- Rae, J. W., Borgeaud, S., Cai, T., Millican, K., Hoffmann, J., Song, F., Aslanides, J., Henderson, S., Ring, R., Young, S., et al. (2021). Scaling language models: Methods, analysis & insights from training gopher. *arXiv preprint arXiv:2112.11446*.
- Sakai, T., Oard, D. W., and Kando, N. (2021). *Evaluating Information Retrieval and Access Tasks: NTCIR’s Legacy of Research Impact*. Springer Nature.
- Schäuble, P. and Sheridan, P. (1998). Cross-language information retrieval (clir) track overview. *NIST SPECIAL PUBLICATION SP*, pages 31–44.
- Thomas, P., Spielman, S., Craswell, N., and Mitra, B. (2023). Large language models can accurately predict searcher preferences. *arXiv preprint arXiv:2309.10621*.
- Vitório, D., Souza, E., Martins, L., da Silva, N. F., de Carvalho, A. C. P. d. L., Oliveira, A. L., and de Andrade, F. E. (2024). Building a relevance feedback corpus for legal information retrieval in the real-case scenario of the brazilian chamber of deputies. *Language Resources and Evaluation*, pages 1–21.
- Wang, L., Yang, N., Huang, X., Jiao, B., Yang, L., Jiang, D., Majumder, R., and Wei, F. (2022). Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*.
- Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, b., Xia, F., Chi, E., Le, Q. V., and Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A., editors, *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.
- Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., and Raffel, C. (2020). mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*.

- Zendel, O., Culpepper, J. S., Scholer, F., and Thomas, P. (2024). Enhancing human annotation: Leveraging large language models and efficient batch processing. In *Proceedings of the 2024 Conference on Human Information Interaction and Retrieval*, pages 340–345.
- Zhang, X., Thakur, N., Ogundepo, O., Kamalloo, E., Alfonso-Hermelo, D., Li, X., Liu, Q., Rezagholizadeh, M., and Lin, J. (2023). Miracl: A multilingual retrieval dataset covering 18 diverse languages. *Transactions of the Association for Computational Linguistics*, 11:1114–1131.

Mineração de Argumentos em Textos de Redes Sociais no Idioma Português

Vitor Domingos Baldoino dos Santos¹, Livia Alabarse dos Santos¹, Orlando B. Coelho (*in memoriam*)¹, Renata Mendes de Araujo^{1,2}, Ivan Carlos Alcântara de Oliveira¹

¹Faculdade de Computação e Informática e Programa de Pós-Graduação em Computação Aplicada - Universidade Presbiteriana Mackenzie
São Paulo - SP - Brasil.

²Programa de Pós-Graduação em Sistemas de Informação - EACH/USP
São Paulo - SP - Brasil

{vitordomingos.santos, liviaalabarse.santos}@mackenzista.com.br
{renata.araujo, orlando.coelho, ivan.oliveira}@mackenzie.br

Abstract. *This paper presents the challenges and research advances aimed at developing computational solutions capable of supporting the understanding of debates on social media in the Portuguese language. One of the fundamental bases of these solutions is the application of Argument Mining techniques. We present the strategies used to address the challenges of argument mining on social media, particularly the use of deep learning. The results obtained show the effectiveness of the selected models for the considered tasks, achieving an F1-Score of 0.85 for sentiment analysis, 0.97 for position detection, and 0.76 for irony detection.*

Resumo. *Este artigo apresenta os desafios e os avanços de pesquisa voltada à construção de soluções computacionais capazes de apoiar o entendimento do debate em redes sociais no idioma português. Uma das bases fundamentais dessas soluções é a aplicação de técnicas de Mineração de Argumentos. Apresentamos as estratégias utilizadas para o endereçamento de desafios da mineração de argumentos em redes sociais, em particular, o uso de deep learning. Os resultados obtidos demonstram boa eficácia dos modelos selecionados para as tarefas consideradas, tendo atingido um F1-Score de 0,85 para a análise de sentimento, 0,97 na detecção de posição e 0,76 na detecção de ironia.*

1. Introdução

A área de Mineração de Argumentação (MA) é uma área multidisciplinar onde se encontram a Linguística Computacional e a Ciência de Dados, cujo objetivo é identificar, extrair e compreender a estrutura de argumentação em textos e/ou discussões [Lawrence e Reed 2020][Lyts et al. 2019][Stede e Schneider 2019].

Argumentação é o processo pelo qual argumentos são construídos, compartilhados e avaliados a partir de outros argumentos [Palau e Moens 2009]. Uma argumentação se estrutura a partir de evidências, premissas, fatos e falas que suportam ou não uma determinada alegação, em uma cadeia de raciocínio que leva à conclusão de discussões e à tomada de decisão [Toulmin 2003]. A argumentação tem papel importante nas atividades humanas, e tem sido compreendida como uma área de pesquisa que surge com base em campos como a Retórica e a Filosofia, mas que hoje

inclui estudos de Processamento de Linguagem Natural e modelos teóricos de discussão [Palau e Moens 2009][Lawrence et. al. 2012], tendo como áreas antecessoras a Mineração de Opiniões e a Análise de Sentimentos, entre outras [Lawrence e Reed 2020][Lytos et.al. 2019].

O projeto HEIWA¹ pretende construir soluções computacionais de análise de redes sociais baseadas em técnicas de MA para a compreensão de discussões em redes sociais, com um olhar específico para o contexto brasileiro. O principal resultado esperado com o projeto é a construção de uma plataforma que permita o acompanhamento de discussões em redes sociais, preferencialmente pelos usuários das próprias redes, mas também para interessados no estudo de comportamentos e mediação em redes sociais. A plataforma será composta por ferramentas computacionais capazes de apoiar um processo de curadoria, mineração de argumentos e visualização do debate para usuários de redes sociais. As implicações da construção dessas tecnologias envolvem aspectos sociais, como o aperfeiçoamento da qualidade do debate e a democracia, aspectos educacionais e de desenvolvimento de pensamento crítico.

Neste artigo, apresentamos os avanços obtidos no escopo do projeto, especificamente em relação ao uso da MA para a identificação de argumentos em textos extraídos de redes sociais em português. Em linhas gerais, os avanços compreendem a necessidade de lidar com linguagem informal, uso de ironia e a polarização presente nos debates online. Além disso, destacamos o uso de modelos de *deep learning*, que se mostraram eficazes em tarefas como análise de sentimento, detecção de posição e detecção de ironia, tendo alcançado bons resultados de *F1-Scores* nas tarefas em um contexto de grande relevância, como os eventos de 8 de janeiro de 2023 em Brasília.

O artigo se estrutura da seguinte forma: a Seção 2 apresenta o conceito e abordagens na literatura para a MA em redes sociais; a Seção 3 descreve as propostas já exploradas no projeto baseadas em análise de sentimentos e *deep learning*; a Seção 4 discute a proposta atual de *pipeline* para execução de tarefas de identificação de tópicos, entidades, sentimentos e ironia; a Seção 5 mostra os resultados dos experimentos computacionais realizados; a Seção 6 conclui o artigo e apresenta os próximos passos da pesquisa.

2. Mineração de Argumentos em Redes Sociais

O processo de identificar, extrair e compreender a estrutura argumentativa a partir de dados textuais é o objetivo da mineração de argumentos (MA) [Stede e Schneider 2019]. A MA tem sido explorada cientificamente para a análise de diferentes conteúdos textuais (documentos, transcrições de debates, transcrições de áudios, etc.), com o objetivo principal de extrair a estrutura de argumentação contida nesses textos e que, eventualmente, levaram para uma deliberação ou decisão [Lawrence e Reed 2020] e para a construção de tecnologias capazes de facilitar discussões ou debater com humanos [Slonim et. al. 2021].

Nos últimos anos, a comunidade científica na área de MA também explora a oportunidade de aplicar o conceito e as técnicas de MA no estudo da argumentação em textos extraídos de redes sociais [Addawood e Bashir 2016][Bosc et. al. 2016][Schaefer

¹ <https://ciberdem.mack.com.br/index.php/projeto-heiwa/>

e Stede 2021][Vecchi et. al. 2021]. Essas pesquisas partem da ideia de que esses textos possam conter uma atividade argumentativa e que a compreensão dos argumentos utilizados durante a interação nas redes sociais pode agregar para a tomada de decisões, gestão da informação e o entendimento do comportamento coletivo em diversos domínios, principalmente as ciências políticas e sociais.

Em um levantamento de literatura preliminar, identificamos que os esforços recentes na aplicação de MA em redes sociais focaram, em alguns casos, na criação e anotação de *datasets* apropriados para a aplicação das técnicas de MA, em maior número, na definição e execução de sequências pré-definidas de tarefas consideradas essenciais para um *pipeline* de mineração de argumentos, e poucos trabalhos na visualização dos resultados (estruturas de argumentação) [Schaefer e Stede 2021].

A despeito dos esforços crescentes, a mineração de argumentos em redes sociais ainda enfrenta desafios significativos. Problemas antigos, apontados desde [Bosc et al. 2016a], permanecem, como a escassez de *datasets* anotados; a baixa qualidade e/ou informalidade dos textos nas redes; o tamanho reduzido das falas (*tweets*, *posts*), que impedem a riqueza de ideias, a ausência de foco ou tópico em discussão, enquanto outros se intensificam, como o crescente custo e restrição de uso das APIs das plataformas de redes sociais de maior escala (*X*, *Facebook* etc). Além disso, a literatura mostra abordagens ainda exploratórias e nenhum dos trabalhos encontrados tratam de conteúdos em português brasileiro.

No que se refere à definição e ao desenvolvimento de *pipelines* para MA em redes sociais, Schaefer e Stede (2021) e Bosc et al. (2016a) mencionam que as tarefas necessárias para aplicação de MA, em linhas gerais, incluem: (i) a anotação de *corpus*; (ii) a detecção de argumentos e suas relações; e (iii) a detecção de posicionamentos. No que se refere à tarefa (i) anotação de *corpus*, a literatura mostra um número escasso de bases anotadas e, pelo menos no material levantado pela presente pesquisa, nenhuma considerando conteúdo em português. No projeto de pesquisa em tela, avançamos nesse aspecto, ao anotar *datasets* de *tweets* em português coletados durante o período eleitoral brasileiro em 2022 [Silva et. al. 2024]. A tarefa seguinte, (ii) detecção de argumentos e relações, envolve, primeiramente, a identificação, no conteúdo de cada postagem, de elementos (alegações), que possam caracterizar as falas como argumentos ou não e, posteriormente, a identificação de relações entre elas (oposição, apoio), permitindo a identificação de estruturas (grafos) de argumentação [Stede e Schneider, 2019][Bosc et. al, 2016a]. Por fim, a tarefa (iii) detecção de posicionamentos consiste em extrair os diferentes pontos de vista sobre um determinado tópico em uma discussão. As tarefas (ii) e (iii) são o foco deste artigo, conforme detalharemos nas seções a seguir.

3. Estudos Iniciais

Os primeiros estudos de MA em redes sociais realizados por esta equipe focaram na identificação de relações de oposição e apoio entre falas em um *corpus* de postagens. As falas dos participantes foram tratadas como sentenças, e as relações entre elas foram identificadas pelo direcionamento das falas entre os interlocutores. Usando os identificadores extraídos das redes sociais, criou-se uma estrutura de dados que preserva a organização entre um argumento e outro, mantendo a estrutura de *threads* de discussão. A natureza da relação (oposição ou apoio) foi determinada pela análise de

sentimentos (negativo, positivo ou neutro).

Os primeiros resultados desta estratégia são reportados por [Sousa et. al. 2021], que explora o pipeline definido por Lippi e Torroni (2016). Para cada uma das sentenças, utilizamos técnicas de análise de sentimentos para determinar a polaridade das manifestações no discurso, utilizando o algoritmo SGD (Pedregosa et al. 2011) para a classificação, treinado com as discussões já classificadas presentes na base *Internet Argument Corpus* (IAC) [Walker et.al. 2012]. A base IAC, contém discussões com um teor politizado, porém amplas o bastante para que pudéssemos usá-la em debates mais genéricos. Quanto ao método de classificação dos argumentos, a classificação se deu sentença a sentença, fornecendo três rótulos que caracterizavam a polaridade de um argumento: apoio, oposição ou neutro. Os resultados obtidos demonstraram uma precisão de 74% na identificação das polaridades dos argumentos, criando assim uma alternativa promissora para a classificação.

Uma segunda abordagem desenvolvida pela equipe envolveu o uso de técnicas de análise de sentimentos baseadas em *deep learning*. A abordagem, apresentada em [Tokuda et. al. 2021], explora a arquitetura BERT [Devlin et al. 2019] como estratégia para elucidar a polaridade das afirmações presentes em um *corpus* de argumentação extraído de uma rede social. Os experimentos realizados demonstraram resultados satisfatórios na extração da polaridade de uma afirmação. A acurácia dos experimentos chegou a 88% em dados não vistos anteriormente pelo modelo, a partir de um conjunto de dados com diversidade extremamente alta de palavras e estruturação livre dos dados, em um formato livre de discussões online com razoável incidência de erros gramaticais e de digitação. Em outro trabalho [Salles e Coelho 2022], foi utilizada uma rede *deep learning* derivada da BERT, a DistilBERT [Zhang et al. 2020], para realizar a análise das emoções presentes em frases que constam do *dataset* em inglês *GoEmotions* [Demszky et al. 2020]. Os resultados obtidos aperfeiçoam os melhores resultados publicados por Cortiz (2021).

Os estudos anteriores demonstram resultados satisfatórios para o uso da análise de sentimentos como parte do processo de MA, mas ainda não avançam em tarefas mais sofisticadas de identificação da estrutura de argumentos e não contemplam conteúdos em português. A estratégia atual do projeto é avançar na identificação de outros componentes de argumentação além de sentimentos, explorando o uso de modelos baseados no *Transformer* [Vaswani et al. 2017] e semelhantes ao BERT aplicados a dados extraídos de redes sociais em português.

4. Proposta de Pipeline para Mineração de Argumentos

Optamos por avançar em nosso projeto a partir da expansão do leque de tarefas consideradas para um *pipeline* de mineração de argumentos, sendo elas: análise de sentimento, identificação de tópico, reconhecimento de entidades, detecção de ironia, detecção de posição, e do uso de redes neurais como principal ferramenta para MA. Destaca-se que as tarefas do *pipeline*, em conjunto, são essenciais para o entendimento do discurso nas redes sociais, visando a posterior identificação de argumentos.

A inclusão das tarefas de identificação de tópico e reconhecimento de entidades nomeadas objetiva desagregar duas grandes atividades vistas na literatura: (i) a

identificação do assunto da discussão; e (ii) a detecção de relação entre pares de texto. Com a desagregação, será possível investigar a eliminação da anotação de um *dataset* para pares de *tweets* relacionados, como o apresentado em [Bosc et al. 2016a]. Adicionalmente, a inclusão da detecção de ironia visa resolver a dificuldade adicional da MA em redes sociais relacionada ao alto grau de informalidade do texto [Schaefer e Stede 2021] e a mudança de semântica gerada pelo uso desse recurso linguístico.

Optamos por não incluir a tarefa de detecção de argumentação, prevista pela literatura, no *pipeline* proposto. Como exposto em [Schaefer e Stede 2021], a literatura recente despendeu muito esforço na construção de *datasets* com alguma definição teórico-conceitual de argumentação. O formato geral do *pipeline* está na Figura 1.

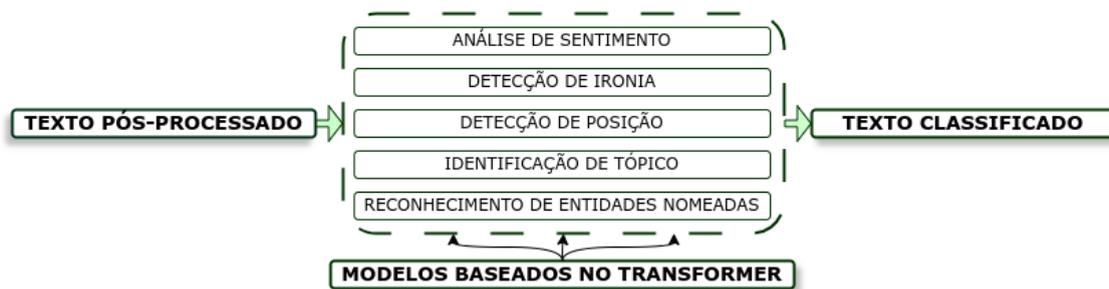


Figura 1. Pipeline proposto da mineração de argumentos.

Posto que o objetivo geral deste trabalho é compreender o debate em uma rede social, cujas características textuais exigem uma análise mais sofisticada de entendimento do texto antes de realizar a identificação da argumentação, entendemos que as tarefas de identificação de tópicos, entidades, sentimentos, posição e ironia de uma discussão oferecem vantagem para o seu entendimento, ainda que ali possamos constatar que não ocorra argumentação. Um exemplo simplificado desta estratégia, tendo por base textos sintéticos sem o pré-processamento e sem considerar todas as tarefas do *pipeline*, é demonstrado na Figura 2. Nela, são apresentados três *tweets* sintéticos e sem pré-processamento a respeito do filme “Duna: Parte Dois”, tópico correspondente a esses *posts*. Com os *tweets* pré-processados, é possível encontrar: as entidades “Duna”, “Aila” e “Paul Atreides”; os sentimentos de cada um deles; e a ironia existente. A partir disso, uma estrutura do debate realizado sobre esse tópico pode ser montada.

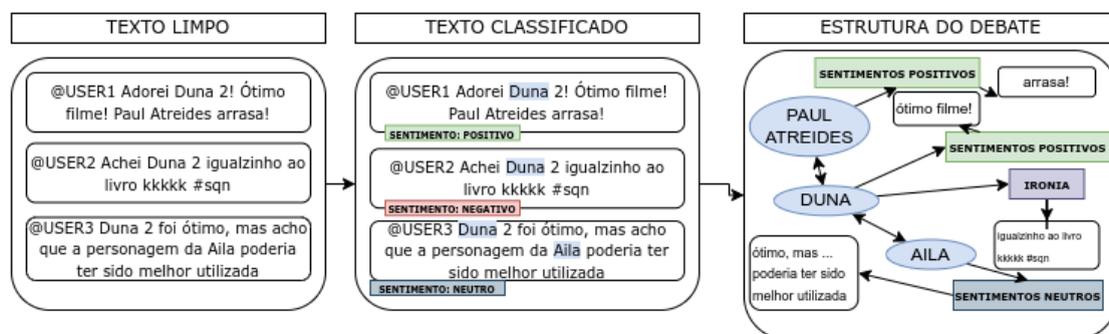


Figura 2. Exemplo do processamento de textos em redes sociais para a identificação da estrutura do debate (entidades, sentimentos e ironia).

5. Experimentos

Para dar cabo das tarefas selecionadas, procurou-se por modelos pré-treinados em português ou que tivessem sofrido *fine-tuning* para tarefas em português com um bom resultado. Um fator crucial para a seleção de modelos foi a sua presença no *HuggingFace Hub*², repositório que visa armazenar e distribuir modelos de *deep learning* para facilitar sua utilização em pesquisas e aplicações comerciais. A Tabela 1 apresenta os modelos selecionados para os experimentos, com a coluna “Modelo” sendo o ID dos modelos na plataforma da *HuggingFace* e coluna ID sendo a utilizada para referenciar os modelos ao longo desta seção.

O modelo 1, apelidado na literatura de BERTimbau (Souza et al., 2020), foi pré-treinado na base de dados brWaC [Wagner Filho et al., 2018], ou seja, em dados sem domínio específico em português brasileiro. Já os modelos 2 e 3, BERTweet.BR [Carneiro, 2023] e BERTaporu [Costa et al., 2023], respectivamente, foram pré-treinados em milhares de *tweets* em português. Entretanto, o modelo 4, RoBERTuito [Pérez et al., 2022], foi pré-treinado em *tweets* na língua espanhola, mas já demonstrou bons resultados para a tarefa de análise de sentimento em português [Pérez et al., 2023].

Tabela 1. Modelos escolhidos para os experimentos computacionais.

ID	Modelo	Endereço Web
1	neuralmind/bert-base-portuguese-cased	huggingface.co/neuralmind/bert-base-portuguese-cased
2	melll-uff/bertweetbr	huggingface.co/melll-uff/bertweetbr
3	pablocosta/bertabaporu-base-uncased	huggingface.co/pablocosta/bertabaporu-base-uncased
4	pysentimiento/robertuito-base-cased	huggingface.co/pysentimiento/robertuito-base-cased

Por sua vez, a base de dados utilizada para realizar o *fine-tuning* dos modelos, [Silva et. al. 2024], foi construída a partir das discussões realizadas em torno dos eventos de 8 de janeiro de 2023 na praça dos Três Poderes, em Brasília, e anotada em etapas manuais e automáticas (com uso de LLMs) para as tarefas de análise de sentimento (AS) (positivo, negativo ou neutro), detecção de ironia (DI) (contém ironia, não contém ironia) e detecção de posição (DP) (a favor da invasão, contra a invasão). Considerando que a base foi construída a partir de um evento marcante, inferimos que o *corpus* já está intrinsecamente associado a um único tópico central. Dessa forma, a tarefa de identificação de tópico foi considerada redundante e removida desta bateria de experimentos. A detecção de posição, por sua vez, foi conduzida em relação a esse tópico pré-definido, permitindo uma análise mais precisa das opiniões expressas no contexto específico do evento. Além disso, optou-se por explorar o reconhecimento de entidades nomeadas (NER) futuramente devido ao tempo significativo que seria necessário para uma anotação manual da base de dados e treinamento do modelo, considerando o cronograma previsto.

A Tabela 2 apresenta a quantidade de *tweets* utilizada para os experimentos de cada tarefa, bem como a distribuição entre as classes de cada tarefa. Deve-se notar que a base está balanceada para a tarefa de AS, mas não para as tarefas de DP e DI. Adicionalmente, cabe destacar que os *tweets* utilizados em cada uma das tarefas não

² <https://huggingface.co/>

possuem total sobreposição, isto é, não são necessariamente os mesmos entre as bases de cada tarefa.

Tabela 2. Quantidade de tweets entre as tarefas e classes.

Tarefa	Classe	Treino	Validação	Teste	Total
Análise de Sentimento	Neutro	234	50	50	334
	Positivo	233	50	51	334
	Negativo	234	50	50	334
Detecção de Posição	Neutro	77	16	17	110
	A favor	19	4	4	27
	Contra	1453	312	312	2077
Detecção de Ironia	Contém ironia	232	50	50	332
	Não contém ironia	615	132	132	879

Os resultados do ajuste dos modelos podem ser vistos na Tabela 3. Os modelos foram treinados com 5 passagens completas pelo conjunto de treino e aquele com melhor desempenho no conjunto de validação foi salvo e utilizado para avaliação no conjunto de teste. Todos os modelos foram treinados variando o número de camadas treinadas (as 2 ou 4 últimas camadas) e o *batch size* (testes realizados com 8, 16, 32 e 64). Utilizou-se a taxa de aprendizado de 0.001 para todos os modelos e a entropia cruzada na função de custo, com ajuste para penalizar mais severamente os erros nas classes minoritárias. Na tarefa de DI, o melhor modelo foi o RoBERTuito (ID 4 - Tabela 1), tendo sido ajustado nas últimas duas camadas e com um *batch size* de 16. Nas tarefas de AS e DP, o melhor modelo foi o BERTimbau (ID 1 - Tabela 1), sendo ajustado nas últimas 4 camadas e com o mesmo *batch size* do modelo anterior.

Tabela 3. Métricas de avaliação dos melhores modelos por tarefa.

Tarefa	Modelo	F1-Score	Precision	Recall
Análise de Sentimento	neuralmind/bert-base-portuguese-cased	0,85	0,86	0,86
Detecção de Ironia	pysentimiento/robertuito-base-uncased	0,76	0,76	0,76
Detecção de Posição	neuralmind/bert-base-portuguese-cased	0,97	0,97	0,97

Embora os resultados obtidos aqui não sejam diretamente comparáveis com a literatura por serem obtidos em uma base de dados diferente, eles são numericamente superiores aos encontrados em Pérez et al. (2023). Comparativamente, o modelo de AS em português ajustado em Pérez et al. (2023), treinado a partir do BERTweet.BR na base de dados apresentada em Brum & Volpe Nunes (2018), alcançou uma macro *F-Score* de 0,73 no conjunto de teste do *dataset* [Silva et. al. 2024]. Além disso, a título de exemplo, a Tabela 4 apresenta dois *tweets* do conjunto de teste, a classificação dada pelo modelo e a classe tida como verdadeira. Os dois *tweets* foram levemente modificados para evitar a sua identificação na rede social sem comprometimento do seu significado e avaliação.

Tabela 4. Exemplos de classificação de dois tweets do conjunto de teste.

Texto	Análise de Sentimento		Detecção de Posição		Detecção de Ironia	
	Classe Verdadeira	Classe Prevista	Classe Verdadeira	Classe Prevista	Classe Verdadeira	Classe Prevista
mano passei a noite toda no meu quarto assistindo tbt e fui ver agr que tava a maior confusao em brasilia k super antenada eu sou	Positivo	Neutro	Contra	Contra	Não contém ironia	Não contém ironia
nem eu acreditei quando vi onde ela tava me mandou corrente de excursao pra brasilia e tudo e nem dei bola	Neutro	Neutro	Contra	Contra	Não contém ironia	Contém ironia

6. Conclusão

Este artigo apresenta os avanços desta pesquisa em andamento voltada à construção de soluções computacionais capazes de apoiar o entendimento do debate em redes sociais. Uma das bases fundamentais destas soluções, explorada neste artigo, é a aplicação de técnicas de mineração de argumentos capazes de identificar a estrutura de argumentação presente nas diversas falas na rede. A expectativa é que a identificação e visualização da estrutura de argumentação possa auxiliar usuários da rede a compreenderem, refletirem e, eventualmente, melhor participarem do debate público.

No treinamento dos modelos, os resultados mais significativos incluem o bom desempenho do modelo BERTimbau na tarefa de detecção de posição, com *F1-Score* de 0,97. Adicionalmente, o modelo RoBERTuito também obteve um desempenho bom, com um *F1-Score* de 0,76 na detecção de ironia, uma tarefa ainda mais complexa no contexto das redes sociais. No entanto, destaca-se como limitação o fato de o *dataset* utilizado ser pequeno e desbalanceado, o que pode comprometer a representatividade estatística necessária para treinar e testar adequadamente modelos de *deep learning*, especialmente na tarefa de detecção de posição. Esse aspecto pode restringir a generalização dos resultados obtidos para diferentes contextos e tópicos. Os resultados de aplicação do *pipeline* proposto nos ajudarão, em passo seguinte, a projetar as abordagens para a realização da tarefa de detecção de argumentação, sofisticando a identificação da estrutura do debate. Como trabalhos futuros, planejamos aumentar a quantidade de dados anotados utilizados para treinamento dos modelos, bem como explorar outras tarefas de um *pipeline* de mineração de argumentos. Outra estratégia prevista é explorar a utilização de *Large Language Models (LLMs)* [Zhao et al. 2023][Brown et al., 2020], abrindo a possibilidade de resumir o *pipeline* para apenas um modelo.

Agradecimentos

Os autores agradecem à FAPESP pelo financiamento desta pesquisa (#2021/14772-1). Renata Araujo é bolsista de produtividade em desenvolvimento tecnológico e extensão inovadora do CNPq (#305645/2022-6). Vitor dos Santos é bolsista TT1 pela FAPESP (#2023/04752-9). Livia Alabarse dos Santos é bolsa TT1 pela FAPESP (2023/04042-1).

Referências

- Addawood, A. e Bashir, M. (2016). “What Is Your Evidence? A Study of Controversial Topics on Social Media”. Em: *Proceedings of the Third Workshop on Argument Mining (ArgMining2016)*. Berlin, Germany. Association for Computational Linguistics. pages 1–11.
- Bosc, T., Cabrio, E. e Villata, S. (2016). “Tweeties Squabbling: Positive and Negative Results in Applying Argument Mining on Social Media”. *Frontiers in Artificial Intelligence and Applications*, v. 287, p. 21–32.
- Bosc, Tom, Cabrio, E. e Villata, S. (2016a). “DART: a Dataset of Arguments and their Relations on Twitter” Em: *Proceedings of the 10th edition of the Language Resources and Evaluation Conference*. pp. 1258-1263.
- Brown, T., Mann, B., Ryder, N., et al. (2020). Language Models are Few-Shot Learners. Em: *Advances in Neural Information Processing Systems*. Curran Associates, Inc.
- Carneiro, F. P. (2023). “BERTweet.BR: A Pre-Trained Language Model for Tweets in Portuguese”. Dissertação de Mestrado. Universidade Federal Fluminense, Programa de Pós-Graduação em Computação. Niterói.
- Cortiz, D. (2021) “Exploring transformers in emotion recognition: a comparison of bert, distillbert, roberta, xlnet and electra”. *arXiv*. arXiv:2104.02041.
- Costa, P. B., Pavan, M. C., Santos, W. R., Silva, S. C., & Paraboni, I. (2023). “BERTabaporu: Assessing a Genre-Specific Language Model for Portuguese NLP”. Em: *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, p. 217–223. Shoumen, Bulgaria. <https://aclanthology.org/2023.ranlp-1.24>
- Demszky, D., Movshovitz-Attias, D., Ko, J., Cowen, A.S., Nemade, G., & Ravi, S. (2020) “GoEmotions: A Dataset of Fine-Grained Emotions”. *arXiv*. arXiv:abs/2005.00547.
- Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K. (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*, Association for Computational Linguistics.
- Lawrence, J., Bex, F., Reed, C. e Snaith, M. (2012) “AIFdb: Infrastructure for the Argument Web.” Em: *Proceedings of the 6th International Conference on Computational Models of Argument*. IOS Press. pp. 515-516.
- Lawrence, J. e Reed, C. (2020) “Argument mining: A survey”. *Computational Linguistics*, v. 45(4), pp. 765-818, 2020.
- Lippi, M., Torroni, P. (2016). “Argumentation mining: State of the art and emerging trends”. *ACM Transactions on Internet Technology*, 16(2), 1-25.
- Palau, R. M. e Moens, M. F. (2009). “Argumentation mining: the detection, classification and structure of arguments in text”. Em: *Proceedings of the 12th International Conference on Artificial Intelligence and Law*. pp. 98-107.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... e

- Duchesnay, E. (2011) "Scikit-learn: Machine learning in Python". *The Journal of machine Learning research*, 12,2825-2830.
- Pérez, J. M., Furman, D. A., Alonso Alemany, L., & Luque, F. M. (2022). "RoBERTuito: A pre-trained language model for social media text in Spanish". Em: *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, p. 7235–7243. European Language Resources Association. <https://aclanthology.org/2022.lrec-1.785>
- Pérez, J. M., Rajngewerc, M., Giudici, J. C., Furman, D. A., Luque, F., Alemany, L. A., & Martínez, M. V. (2023). "pysentimiento: A Python Toolkit for Opinion Mining and Social NLP tasks". arXiv. <http://arxiv.org/abs/2106>.
- Salles, G. T., Coelho, O. B. (2022). "Reconhecimento de Emoções em Mineração de Argumentos com Deep Learning". Trabalho de Conclusão de Curso. Universidade Presbiteriana Mackenzie.
- Schaefer, R. e Stede, M. (2021). "Argument Mining on Twitter: A survey". *Information Technology*, v. 63, n. 1, p. 45–58.
- Silva, L. J., Santos, L. A.; Araujo, R., Coelho, O. B., Correa, A. G, D.; Oliveira, I. C. A. (2024) "Tweet_Eleicoes_2022: Um dataset de tweets durante as eleições presidenciais brasileiras de 2022". Brazilian Workshop on Social Network Analysis and Mining (BRASNAM), 13. Brasília/DF. Porto Alegre: Sociedade Brasileira de Computação. p. 193-199. <https://doi.org/10.5753/brasnam.2024.1940>.
- Slonim, N., Bilu, Y., Alzate, C., Bar-Haim, R., Bogin, B., Bonin, F., ... e Aharonov, R. (2021). "An autonomous debating system". *Nature*, 591(7850), p. 379-384.
- Sousa, J.P.S., Nascimento, R. C. U., Araujo, R. M., Coelho, O. B. (2021). "Não se perca no debate! Mineração de Argumentação em Redes Sociais". Brazilian Workshop on Social Network Analysis and Mining (BRASNAM). Porto Alegre: Sociedade Brasileira de Computação. p. 139-150. <https://doi.org/10.5753/brasnam.2021.16132>.
- Souza, F., Nogueira, R., & Lotufo, R. (2020). "BERTimbau: Pretrained BERT Models for Brazilian Portuguese", p. 403–417. https://doi.org/10.1007/978-3-030-61377-8_28
- Stede, M. e Schneider, J. (2019). "Argumentation Mining". Springer. Synthesis Lectures on Human Language Technologies.
- Sun, C., Qiu, X., Xu, Y. e Huang, X. (2019). "How to Fine-Tune BERT for Text Classification?" In Chinese Computational Linguistics. Lecture Notes in Computer Science. Springer International Publishing.
- Tokuda, N. H., Coelho, O. B., Araujo, R.M. (2021). "Análise de Sentimento por meio de Deep Learning aplicada à Mineração de Argumentos". Trabalho de Conclusão de Curso. Universidade Presbiteriana Mackenzie.
- Toulmin, S. E. (2003). *The uses of argument*. Cambridge University Press.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, T. e Polosukhin, I. (2017). "Attention is All you Need". Em: *Advances in Neural Information Processing Systems*. Curran Associates, Inc. 30.
- Vecchi, E. M., Falk, N., Jundi, I., Lapesa, G. (2021). "Towards Argument Mining for

- Social Good: A Survey”. Em: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*. Online. Association for Computational Linguistics. p. 1338–1352.
- Wagner Filho, J. A., Wilkens, R., Idiart, M., & Villavicencio, A. (2018). "The brWaC Corpus: A New Open Resource for Brazilian Portuguese". Em: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association (ELRA). <https://aclanthology.org/L18-1686>
- Walker, M. A., Tree, J. E. F., Anand, P., Abbott, R. e King, J. (2012). “A Corpus for Research on Deliberation and Debate”. Em: *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC '12)* v. 12. Istanbul, Turkey. p. 812–817.
- Zhang, T., Wu, F., Katiyar, A., Weinberger, K. Q., & Artzi, Y. (2020) “Revisiting few-sample BERT fine-tuning”. arXiv preprint arXiv:2006.05987.
- Zhao, W. X., Zhou, K., Li, J., et al. (2023). “A Survey of Large Language Models”. Arxiv. arXiv. <http://arxiv.org/abs/2303.18223>.

A Hybrid Machine Learning Method to Author Name Disambiguation

Natan S. Rodrigues^{1,2}, Celia G. Ralha²

¹Instituto Acadêmico de Ciências Tecnológicas
Universidade Estadual de Goiás (UEG) – Goiás – GO – Brasil

²Departamento de Ciência da Computação – Instituto de Ciências Exatas
Universidade de Brasília (UnB) – Brasília, DF – Brasil

natan.rodrigues@ueg.br, ghedini@unb.br

Abstract. *Digital bibliographic repositories, including publications, authors, and research fields are essential for sharing scientific information. Nevertheless, the information retrieval, extraction, and classification efficiency in such archives is threatened by author name ambiguity. This paper addresses the Author Name Disambiguation (AND) problem by proposing a hybrid machine learning method integrating Bidirectional Encoder Representations from Transformers (BERT), Graph Convolutional Network (GCN), and Graph Enhanced Hierarchical Agglomerative Clustering (GHAC) approaches. The BERT model extracts textual data from scientific documents, the GCN structures global data from academic graphs, and GHAC considers heterogeneous networks' global context to identify scientific collaboration patterns. We compare the hybrid method with AND state-of-the-art work using a publicly accessible data set consisting of 7,886 documents, 137 unique authors, and 14 groups of ambiguous authors, along with recognized validation metrics. The results achieved a high precision score of 93.8%, recall of 96.3%, F1-measure of 95%, Average Cluster Purity (ACP) of 96.5%, Average Author Purity (AAP) of 97.4% and K-Metric of 96.9%. Compared to the AND baseline approach, the hybrid method presents better results indicating a promising approach.*

1. Introduction

Digital bibliographic repositories are vast reservoirs of bibliographic citation information (DBLP [DBLP 2024], ArnetMiner [AMiner 2024b], CiteSeerX [CiteSeerX 2019]). They offer functionalities that allow the identification of works by scientists, authors, and their respective academic social networks. The DBLP currently lists around 7 million works in Computer Science, including journals and conference articles. In January 2024, DBLP gathered information on approximately 3.5 million authors, with 227 thousand names of researchers and publications manually verified by the DBLP team, corresponding to a curation of 34% of all publications in the database.¹ ArnetMiner stores information on approximately 2 million of scientific works, 1.7 million of authors, and 8 million of bibliographic citations [AMiner 2024a].²

By storing millions of information from bibliographic records, digital repositories become an essential source of information for the global academic and scientific

¹<https://dblp.org/>

²<https://www.aminer.org/>

community, allowing retrieval, extraction and classification of relevant publications in a centralized manner [Ferreira et al. 2020]. In addition to these bibliographic features, such digital libraries provide helpful analysis useful for better decision-making by scientific funding agencies and academic institutions [Hussain and Asghar 2017].

However, a common problem in digital bibliographic repositories is automatic Author Name Disambiguation (AND). The AND problem occurs when different authors have the same name record or when an author has multiple name records in the same data set. Such a problem can significantly affect the document and information retrieval performance through *Web* search engines and obstruct entity integrity for integrated databases. Even though the author's name ambiguity problem has been studied for decades, it remains without a canonical solution. Thus, research efforts to solve the AND problem are essential, especially considering that digital bibliographic repositories are becoming more person-centric than document-centric [Shin et al. 2014].

This work addresses the AND problem with a novel hybrid method combining advanced machine learning techniques, such as the Bidirectional Encoder Representations from Transformers (BERT) [Devlin et al. 2019], Graph Convolutional Network (GCN) [Zhang et al. 2019], and Graph Enhanced Hierarchical Agglomerative Clustering (GHAC) [Qiao et al. 2019]. As proposed by [Kipf and Welling 2017], GCN is a powerful machine learning model that extends the Convolutional Neural Network (CNN) to handle data structured as graphs, capturing local and global dependencies within a network. Our method aims to enhance the AND accuracy in digital bibliographic repositories considering information retrieval, extraction, and classification by applying document content semantic treatment related to a graph representation of relationships between documents, authors, and other scientific attributes.

As presented in [Ferreira et al. 2020], there are several approaches to solving AND problem applying various techniques, but no works combining transfer learning with GCN and GHAC techniques. Also, according to a recent AND literature review, using the theory of the consolidated meta-analytic approach with quantitative techniques and bibliometric aspects, the hybrid method proposed is considered a novel solution to AND [Rodrigues et al. 2024].

The rest of the article includes in Section 2 related work focusing on approaches to the AND problem. Section 3 details the AND hybrid method. Section 4 includes the conducted experiments with the evaluation metrics. In Section 5, we present the results with discussion. Finally, the conclusion and future work are in Section 6.

2. Related Work

As presented in [Rodrigues et al. 2024], largely used AND solving approaches are author grouping associated with similarity functions and clustering methods, and some works with author assignment allied to classification methods. Also, approaches based on graphs, word embedding with supervised learning, and heuristics with probabilistic applications are common. The literature review highlights author clustering techniques' prevalence and effectiveness, especially when addressing issues associated with large bibliographic databases. In this section, we present works most related to our hybrid method.

The authors in [Kim and Owen-Smith 2020] explore supervised techniques using transfer learning on AND tasks where no labelled data was available for training. The

results show that by training source data that well represent the main characteristics of the target datasets, the developed disambiguation models through transfer learning can produce results comparable to those achieved by traditional machine learning approaches, which train algorithms on specifically labelled subsets of the target data.

In [Waqas and Qadir 2021], the authors propose a method to perform AND based on heuristic clusters in several layers. They used global characteristics and those related to the structure of publications to group them. One of the differences pointed out by the authors is that instead of relying only on keyword information, the approach also considers the contextual structure of publications for grouping. The authors use an incremental classification method to reduce errors after creating clusters. A dataset called *CustAND* was presented for testing and executing the AND method.

The approach of [Pooja et al. 2022] uses GCN in conjunction with attention mechanisms for learning representations in a heterogeneous graph of documents. The work highlights the importance of using attention at different levels, both about the types of neighbors and relationships, to incorporate relevant context into learning node representations. The emphasis on attention allows a detailed analysis of the impact of this mechanism on capturing semantic and contextual information from documents in a graph. The authors used two ArnetMiner variants as data sets, the first with 110 and the second with 100 ambiguous name references.

3. The Hybrid Method

The hybrid method has four main steps as presented in Figure 1 and described in the sequence.

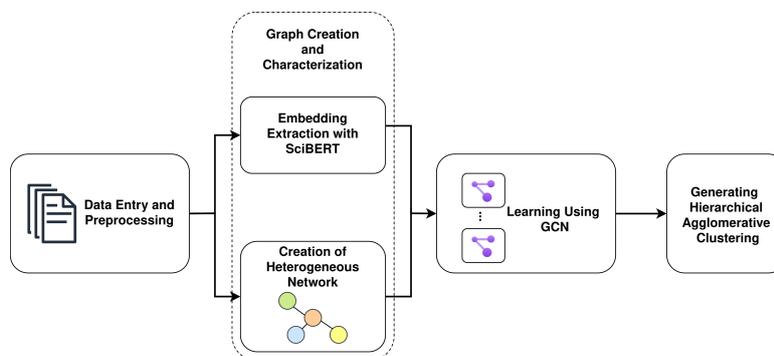


Figure 1. The hybrid method workflow.

3.1. Data Entry and Preprocessing

The hybrid method's first step deals with the input and preprocessing of document data (publications), when data is adjusted and formatted to ensure input suitability for the subsequent steps.

3.2. Graph Creation and Characterization

This step plays a fundamental role as it creates the structure of the heterogeneous network from the information received from the previous step and provides contextual information for the AND task.

- **Creation of a Heterogeneous Network** - includes different types of nodes and edges. The graph is formally defined as $G_{\text{heterogeneous}} = (N_{\text{nodes}}, E_{\text{edges}})$, where N_{nodes} are nodes representing publications, authors, and words. The edges E_{edges} represent the different connections between nodes, such as *contains* (between publications and words), *written_by* (between publications and authors), *co-authored* (between authors who collaborated on the same publication), and *shared-word* (between publications that share keywords).
- **Embedding Extraction with BERT** - BERT uses transfer learning pretraining its parameters on large sets of unlabeled texts with only minor modifications to perform tasks in a given domain. BERT converts every word in a text into a vector representation that captures the word's meaning given the context in which it appears. This representation can be combined to obtain a representation of entire sentences. In this work, we use the SciBERT variant of BERT pre-trained on scientific texts, which is particularly effective at capturing the contextual and semantic information of academic documents [Beltagy et al. 2019]. SciBERT calculates the embeddings of publications based on titles and abstracts. These embeddings are incorporated as features of the nodes that represent the publications in the graph. The algorithm is described in the sequence.

Given a set of N documents with titles and abstracts, where each document i has a title T_i and an abstract R_i , the embedding extraction process with SciBERT can be detailed as follows:

1. **Tokenization of Titles and Abstracts:** each title T_i and abstract R_i are tokenized into sequences of separate tokens, represented by $\{t_{i,1}, t_{i,2}, \dots, t_{i,L_i}\}$ and $\{r_{i,1}, r_{i,2}, \dots, r_{i,M_i}\}$, respectively.
2. **SciBERT Embedding Generation:** the token sequences of the titles T_i and abstracts R_i are processed by BERT, which produces a vector of embeddings for each document. These embeddings capture the semantics of the texts, reflecting the main topics and contextual relations.
3. **Graph Embedding:** the embeddings resulting from SciBERT are used as nodes features that represent the publications in the heterogeneous graph.

In our algorithm, SciBERT performs the embedding extraction on the titles and abstracts of the documents. These embeddings are used as node features in the heterogeneous graph, allowing the subsequent GCN to use these representations to analyze the interactions between publications, authors, and words, including co-authorship relationships. An edge index represents sparsely the connections between nodes. This format allows the GCN to process large heterogeneous networks while maintaining essential connectivity information among entities.

3.3. Learning Using GCN

After extracting embeddings with SciBERT and constructing the heterogeneous graph, the titles and abstract embeddings are used as features of the nodes in the network. The propagation operations in the GCN layers use these embeddings to compute representations of neighboring nodes. The need to apply a GCN model to a heterogeneous network, instead of other traditional deep learning techniques arises from the particularities of networks, where relationships between different types of nodes and graph structures must be captured effectively.

In the GCN step, this proposed hybrid method initially processes the textual data to create a vocabulary with a feature matrix, where each row corresponds to the embedding of a node, such as documents, authors, or keywords. The edge index represents the connections between nodes, preserving the essential relationships in the heterogeneous network.

To capture local and global dependencies within the heterogeneous network data, each layer of the GCN updates the node representations based on the connections and features defined by the edge index. The proposed GCN model uses activation functions to introduce nonlinearities in the model. In our work, we use the ReLU function ($\sigma(x) = \max(0, x)$) at different stages of GCN (widely used to mitigate the vanishing gradient problem). GCN training is performed by minimizing the MSE loss function, defined as $\mathcal{L} = \frac{1}{N} \sum_{i=1}^N (\mathbf{Z}_i - \mathbf{X}_i)^2$, where N is the number of nodes in the network, \mathbf{Z}_i the final GCN output for node i , and \mathbf{X}_i the original feature vector of node i . The Adam optimizer [Kingma 2014] was used to adjust the model’s weights, adjusting an initial learning rate as needed.

Finally, GCN produces embeddings of the nodes in the network, represented as low-dimensional vectors that capture both the nodes’ initial features and the network structure. These embeddings are used for subsequent tasks, such as agglomerative hierarchical clustering, which will be performed in the next step.

3.4. Generating Hierarchical Agglomerative Clustering

The disambiguated authors’ clustering results are generated based on their representations in the heterogeneous network. The goal is to group documents with similar characteristics the interactions between publications, authors, and keywords using the GHAC method [Qiao et al. 2019]. The GHAC is an agglomerative hierarchical clustering algorithm that integrates network structural information considering the average similarity of the embeddings between the connected nodes. The algorithm is suitable for complex heterogeneous networks as the one built from the embeddings generated by GCN.

Initially, each document is an individual cluster. The iterative algorithm proceeds to merge the clusters with the highest average similarity between their components until reaching the desired number of clusters. The similarity between the two clusters is defined based on the normalized inner products of the node embeddings, allowing GHAC to capture the semantic and contextual data relationships.

Documents are grouped to maximize the internal cohesion of the clusters while preserving the semantic and structural interaction characteristics between the different types of entities in the network. This method not only groups documents based on local similarities but also considers the global context of the heterogeneous network, making it particularly effective in organizing complex academic networks and identifying underlying co-authorship and scientific collaboration patterns.

4. Experiments

To validate the hybrid method, we conducted experiments comparing to the multi-layer approach with clustering techniques of [Waqas and Qadir 2021] as a baseline, using the public data set *CustAND*,³ which is composed of 14 ambiguous name groups with 137

³https://github.com/humaira699/CustAND_Full.git

distinct authors and 7,886 documents [Waqas and Qadir 2022]. This dataset is valuable for AND studies with various attributes and complex data relationships. The execution pipeline and the code for implementing this method are available in the repository.⁴

4.1. Experimental Setup

We used the document titles and abstracts to extract embeddings with SciBERT. We then concatenated these features to form the input text tokenized using the BERT tokenizer limited to 512 tokens. The output was a 768-dimensional embedding representing each document. The empirically defined GCN configuration includes three layers with an embedding size of 768, ReLU activation function, Mean Squared Error (MSE) loss function, and the optimization performed with a 0.001 learning rate for the Adam algorithm. We executed the training for 200 epochs with a batch size of 128. Python language was used to execute the experiments in a Google Colab L4 environment with the hardware accelerator L4, GPU with 22.5 GB of RAM, CPU with 53 GB of RAM, 201.2 GB disk, and the runtime type configured for Python 3.

4.2. Evaluation Metrics

The precision, recall, F1-measure, and specific metrics for clustering, such as Average Cluster Purity (ACP), Average Author Purity (AAP), and K-Metric metrics commonly presented in the AND literature are used to evaluate the experimental results.⁵

Precision measures the proportion of correctly classified documents relative to the total number of author documents, assessing the algorithm’s ability to assign documents to authors correctly as $\text{Precision} = \frac{\text{Documents Correctly Classified}}{\text{Total Documents Classified}}$. Recall evaluates the ability of the algorithm to retrieve all documents from a real author, measuring the retrieval capacity of the algorithm about real authors as $\text{Recall} = \frac{\text{Documents Correctly Retrieved}}{\text{Total Documents from Real Author}}$. F1-measure is the harmonic mean of precision and recall, providing a balanced metric between these metrics (general performance metric) as $\text{F1-measure} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$.

ACP evaluates the average purity of the clusters generated by the algorithm about the theoretical clusters. ACP measures how well the documents were grouped into clusters that represent real authors as $\text{ACP} = \frac{1}{N} \sum_{i=1}^q \sum_{j=1}^R \frac{n_{ij}^2}{n_i}$, where N is the total size of the publication/paper records in the test set, q is the number of hybrid method/predicted clusters, R is the number of manually generated reference/real clusters, n_{ij} is the number of elements in common between the hybrid method-predicted clusters i and the reference clusters j , and n_j is the number of elements in the reference cluster j . The purer the clusters, the higher the ACP value. AAP measures how fragmented or cohesive the clusters predicted by the algorithm are relative to the reference clusters. A higher AAP indicates that the clusters are less fragmented, as $\text{AAP} = \frac{1}{N} \sum_{j=1}^R \sum_{i=1}^q \frac{n_{ij}^2}{n_j}$.

K-metric determines the trade-off between the average purity of clusters (ACP) and the average purity of authors (AAP). It is a metric that provides a single measure that considers both the quality of clusters and the quality of document attribution to real authors, as $\text{K-metric} = \sqrt{\text{ACP} \times \text{AAP}}$. K-metric helps evaluate the overall performance

⁴https://github.com/natansr/adan_hybrid_method.git

⁵Cluster purity measures how well the items in a cluster belong to the same real class. For AND it reflects the authorship records belonging to a single author within a cluster. A higher purity indicates a more homogeneous cluster where one is the ideal value [Ferreira et al. 2020].

of the disambiguation algorithm by balancing the quality of clusters and the quality of document attribution.

Figure 2 presents an illustrative example with geometric figures corresponding to an authorship record, where equal figures represent the same author. There are three theoretical clusters and four empirical ones, with one empirical cluster not pure and two authorship circle records fragmented across two clusters. The results of the metrics applied to this example, considering the ACP with the empirical clusters include in the first two clusters three author records ($\frac{3^2}{3}$), the third and fourth clusters two different authors ($\frac{1^2}{2}$), and the last cluster has a single record ($\frac{1^2}{1}$), the ACP is 0.888 ($\frac{1}{9} \times (\frac{3^2}{3} + \frac{3^2}{3} + \frac{1^2}{2} + \frac{1^2}{2} + \frac{1^2}{1})$). The AAP values numerators remain the same, but the denominators reflect the number of records in the theoretical clusters. For instance, $\frac{3^2}{4}$ represents three records from the same author in an empirical cluster out of four in the theoretical one. The final AAP value is 0.722 ($\frac{1}{9} \times (\frac{3^2}{4} + \frac{3^2}{3} + \frac{1^2}{4} + \frac{1^2}{2} + \frac{1^2}{2})$), and the K-metric is the geometric mean of ACP and AAP ($\sqrt{0.888 \times 0.722} = 0.8$). Precision is 0.857 considering the sum of three authorship record pairs from the same author in the first and second empirical clusters and none in the last three clusters. The denominator sums the total number of authorship record pairs from each empirical ($\frac{3+3+0+0+0}{3+3+1+0}$). Recall is 0.6 using the same Precision numerator with the denominator the sum of the authorship record pairs that refer to the same author in the theoretical clusters 6, 3, and 1 in the first, second, and third theoretical clusters, respectively ($\frac{3+3+0+0+0}{6+3+1}$). Finally, the F1-measure = $\frac{2 \times (0.857 \times 0.6)}{0.857 + 0.6} = 0.7$.

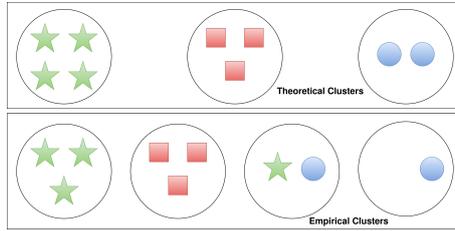


Figure 2. Theoretical and empirical clusters.

5. Results and Discussion

In this section, we present our hybrid method results with the evaluation metrics (Section 4.2) for the *CustAND* dataset with 14 groups of ambiguous names compared to the baseline work of [Waqas and Qadir 2021]. In the *CustAND* dataset, an example of an ambiguous name group for “A Choudhary” consists of 12 distinct authors that share the same name in document citation, namely “Ashish Choudhary”, “Amit Choudhary”, “Anil Choudhary”, “Arvind Choudhary”, “Anupam Choudhary”, “Ajay Choudhary”, “Abhishek Choudhary”, “Aniruddha Choudhary”, “Anjali Choudhary”, “Arjun Choudhary”, “Akshay Choudhary”, and “Arun Choudhary”. Table 1 summarizes the metrics for each ambiguous name group presenting average values for our method and the baseline.

Analysis of our method performance metrics for the 14 ambiguous name groups of the *CustAND* dataset reveals attractive results. Compared to the results reported by [Waqas and Qadir 2021], the average precision across the 14 groups is slightly lower (93.8% versus 94.6%), which may indicate a loss of precision when classifying documents for specific authors. However, the higher Recall (96.3% versus 92.5%) suggests

that the method applied to ambiguous groups has a better recall capacity and is more efficient in identifying all documents of an author. The F1-measure of 95% across the 14 groups, compared to 93.5% for [Waqas and Qadir 2021], demonstrates that the method achieves a better balance between Precision and Recall.

The ACP and AAP metrics across the 14 groups also outperform the baseline with values of 96.5% and 97.4%, compared to 95.8% and 87%, respectively. These results suggest a higher average purity of the generated clusters and a lower fragmentation of the predicted clusters, reflecting a more cohesive and representative grouping of the real authors. Finally, the 96.9% K-metric in the 14 clusters of our method is significantly higher than the 91.24% reported by [Waqas and Qadir 2021], indicating that our method achieves a superior balance between the quality of the clusters and the correct attribution of documents to authors.

Table 1. Performance metrics by ambiguous name group.

Ambiguous Name Group	# Authors	Precision	Recall	F1-measure	ACP	AAP	K-metric
A Choudhary	12	1.000	1.000	1.000	1.000	1.000	1.000
J Martin	9	1.000	1.000	1.000	1.000	1.000	1.000
M A Qadir	15	1.000	1.000	1.000	1.000	1.000	1.000
J Mitchell	10	1.000	1.000	1.000	1.000	1.000	1.000
A Gupta	8	0.853	0.878	0.865	0.875	0.875	0.875
J Robinson	12	1.000	1.000	1.000	1.000	1.000	1.000
A Kumar	9	1.000	1.000	1.000	1.000	1.000	1.000
J Smith	12	0.938	0.988	0.964	0.972	0.972	0.972
Bin Li	8	0.592	0.671	0.632	0.763	0.889	0.826
S Kim	10	0.754	0.944	0.839	0.897	0.895	0.896
D Eppstein	3	1.000	1.000	1.000	1.000	1.000	1.000
Z Zhang	10	1.000	1.000	1.000	1.000	1.000	1.000
J Lee	8	1.000	1.000	1.000	1.000	1.000	1.000
K Tanaka	11	1.000	1.000	1.000	1.000	1.000	1.000
Baseline [Waqas and Qadir 2021]	137	0.946	0.925	0.935	0.958	0.870	0.912
Our Method	137	0.938	0.963	0.950	0.965	0.974	0.969

6. Conclusion

The main objective of this work was accomplished by proposing and evaluating the resolution capacity of the AND problem using a hybrid method that involves transfer learning with SciBERT, GCN, and GHAC. When comparing the effectiveness of our hybrid method with the state-of-the-art work of [Waqas and Qadir 2021], using the *CustAND* dataset, we note that the proposed method outperformed the baseline regarding average accuracy, considering five of six commonly used metrics of precision, recall, F1-measure, ACP, AAP, and K-metric.

Future experiments include comparison to [Pooja et al. 2022] including the use of other machine learning methods, diverse textual extract information methods, and the adoption of graph neural networks approaches, such as Graph Attention Network (GAT) and GraphSAGE with larger datasets. Also, a manageable data entry implementation for the end user as a graphical user interface to make the solution more user-friendly.

References

- AMiner (2005-2024b). Search and mining of academic social networks. <https://www.aminer.org/>. Tsinghua University, Beijing, 100084. China.
- AMiner (2024a). Aminer dataset. Disponível em <https://www.aminer.cn/data/?nav=openData>.
- Beltagy, I., Cohan, A., and Lo, K. (2019). Scibert: A pretrained language model for scientific text. *CoRR*, abs/1903.10676.
- CiteSeerX (2007-2019). Scientific literature digital library and search engine. <https://citeseerx.ist.psu.edu/index>. Pennsylvania State University, University Park, PA 16802, USA.
- DBLP (1993-2024). The digital bibliography & library project. <https://dblp.uni-trier.de/>. Schloss Dagstuhl, Leibniz-Zentrum für Informatik, LZI GmbH.
- Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. page 4171–4186, Minneapolis, Minnesota, USA. Proceedings of NAACL-HLT 2019, Association for Computational Linguistics.
- Ferreira, A. A., Gonçalves, M. A., and Laender, A. H. F. (2020). Automatic disambiguation of author names in bibliographic repositories. *Synthesis Lectures on Information Concepts, Retrieval, and Services*, 12(1):1–146.
- Hussain, I. and Asghar, S. (2017). A survey of author name disambiguation techniques: 2010-2016. *Knowledge Eng. Review*, 32:e22.
- Kim, J. and Owen-Smith, J. (2020). Model reuse in machine learning for author name disambiguation: An exploration of transfer learning. *IEEE Access*, 8:188378–188389.
- Kingma, D. P. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.
- Kipf, T. N. and Welling, M. (2017). Semi-supervised classification with graph convolutional networks. *International Conference on Learning Representations (ICLR)*.
- Pooja, K. M., Mondal, S., and Chandra, J. (2022). Exploiting higher order multi-dimensional relationships with self-attention for author name disambiguation. *ACM Transactions on Knowledge Discovery from Data*, 16(5).
- Qiao, Z., Du, Y., Fu, Y., Wang, P., and Zhou, Y. (2019). Unsupervised author disambiguation using heterogeneous graph convolutional network embedding. In *2019 IEEE International Conference on Big Data (Big Data)*, pages 910–919.
- Rodrigues, N. S., Mariano, A. M., and Ralha, C. G. (2024). Author name disambiguation literature review with consolidated meta-analytic approach. *International Journal on Digital Libraries*, pages 1–21.
- Shin, D., Kim, T., Choi, J., and Kim, J. (2014). Author name disambiguation using a graph model with node splitting and merging based on bibliographic information. *Scientometrics*, 100(1):15–50.
- Waqas, H. and Qadir, A. (2022). Completing features for author name disambiguation (AND): An empirical analysis. *Scientometrics*, 127(2):1039–1063.

- Waqas, H. and Qadir, M. A. (2021). Multilayer heuristics based clustering framework (MHCF) for author name disambiguation. *Scientometrics*, 126(9):7637–7678.
- Zhang, S., Tong, H., Xu, J., and Maciejewski, R. (2019). Graph convolutional networks: a comprehensive review. *Computational Social Networks*, 6(11).

Adapting LLMs to New Domains: A Comparative Study of Fine-Tuning and RAG strategies for Portuguese QA Tasks

Leandro Yamachita da Costa¹, João Baptista de Oliveira e Souza Filho¹

¹Programa de Engenharia Elétrica

Universidade Federal do Rio de Janeiro (UFRJ) – Rio de Janeiro, RJ – Brazil

leandro.yamachita@coppe.ufrj.br, jbfilho@poli.ufrj.br

***Abstract.** The rise of Large Language Models (LLMs) represented a significant advance in text generation applications. However, LLMs face challenges in domains outside the scope of their original training. This study investigates the following two approaches to adapt LLMs to new domains in the context of generative question-answering (QA) with data in Portuguese: fine-tuning and Retrieval-Augmented Generation (RAG). The experiments carried out in this study demonstrate the effectiveness of incorporating external data sources, even in models that had not been adjusted for the specific domain. Furthermore, the combination of supervised fine-tuning with RAG proved to be the most effective approach.*

1. Introduction

The rise of Large Language Models (LLMs) marked a significant advance in Natural Language Processing (NLP), especially in text generation tasks [Brown et al., 2020; Achiam et al., 2023]. These models, which are trained with large volumes of data, can retain vast amounts of knowledge implicitly in their parameters. However, LLMs face challenges in domains outside the scope of their original training data, such as areas of specialized knowledge or current affairs [Kandpal et al., 2023, Kasai et al., 2024]. This issue accentuates the need to adapt LLMs to specific contexts, especially for smaller models with limited memory capacity.

This study explores the adaptation of LLMs to new domains in the generative question-answer (QA) task, a scenario where the model generates answers based on questions provided to it. Fine-tuning is a common approach to adjust LLMs to new domains by modifying the model parameters with training on application-specific data. In the QA task, fine-tuning can be done with question-answers pairs in a "closed-book" scenario [Zhang et al., 2024], where the model does not have access to external information. Nonetheless, this approach may require considerable computational power and extensive data annotation work [Guo et al., 2023].

A widely adopted alternative is the integration of external knowledge sources, such as documents and books, in a setting known as "open-book" [Zhang et al., 2024]. A typical strategy of this approach is the Retrieval-Augmented Generation

(RAG) [Lewis et al., 2020], which combines an information retriever (IR) model, aimed at searching relevant data on an external data source, with a language model to generate answers based on the information retrieved. This strategy allows LLMs to adapt to new domains without the need for fine-tuning.

This work analyzed these two approaches to adapt LLMs to specific domains in the QA task with data in Portuguese. We analyzed both fine-tuning and RAG configurations, in addition to the integration of the two. The experiments demonstrated the effectiveness of incorporating external data sources for improved results. Moreover, different fine-tuning strategies have shown to be particularly effective when combined with the inclusion of external data, even with a reduced volume of training data. For all analysis, we considered a scenario with limited computational resources, where we only used a general-purpose GPU (Nvidia GeForce RTX 4090). The fine-tuning of the models was performed using the QLoRA technique [Dettmers et al., 2023] with quantized models.

2. Background and Related Work

In this session we briefly discuss RAG and LLMs fine-tuning.

2.1. Retrieval-augmented Language Models

Conditioning LLM responses to information from external sources has proven effective in adapting these models to specific domains in several NLP tasks. The RAG approach has been successfully applied in areas such as agriculture [Balaguer et al., 2024], scientific literature [Lála et al., 2023], and medical data [Zakka et al., 2024]; attesting its utility in improving the accuracy and relevance of answers. Additionally, it can reduce the occurrence of hallucinations [Borgeaud et al., 2022; Shuster et al., 2021], improve the model's ability to manage the gradual decline in its knowledge over time [Vu et al., 2023], and enhance the interpretability of the answers [Lewis et al., 2020; Izacard et al., 2023].

The effectiveness of RAG depends on the quality of the retrieval mechanism used, which impacts the relevance of the contextual information obtained. Among traditional retrieval mechanisms, those based on term frequency stand out, which employ sparse representations of text passages, such as TF-IDF [Sparck Jones, 1972] and BM25 [Robertson and Zaragoza, 2009]. Alternatively, more recent approaches employ dense representations of texts, such as Dense Passage Retriever (DPR) [Karpukhin et al., 2020] and ColBERT [Khattab and Zaharia, 2020].

2.2. Fine-tuning

To adapt QA models to a new domain, fine-tuning seeks to adjust the model to respond according to the pattern observed in the training data. Furthermore, it is expected that with fine-tuning the model will acquire domain-specific knowledge, enhancing its capacity to provide more accurate answers.

Fine-tuning large language models (LLMs) typically requires significant computational resources. Parameter Efficient Fine Tuning (PEFT) [Xu et al., 2023] addresses this by freezing the model's parameters and adjusting only the newly added ones. Among PEFT methods, Low-Rank Adaptation (LoRA) [Hu et al., 2021] reduces the number of trainable parameters using low-rank matrices. The Quantized Low-Rank Adaptation (QLoRA) [Dettmers et al., 2023] takes this approach a step further by applying LoRA to quantized models. Studies indicate that PEFT-tuned models often perform comparably to those fully fine-tuned [Li et al., 2023].

2.3. RAG and Fine-tuning

The effectiveness of RAG and fine-tuning strategies has been extensively studied. [Balaguer et al., 2024] compare these methods in a QA model for agriculture, while [Ovadia et al., 2023] extend the comparison to various topics with a multiple-choice QA model. Many RAG models undergo fine-tuning, such as [Lewis et al., 2020], where both the model and retriever are adjusted together. [Zhang et al., 2024] introduce RAFT (Retrieval-Augmented Fine-Tuning), which helps the model ignore irrelevant documents. Overall, pre-trained models require additional fine-tuning to learn specific reading comprehension tasks, which is essential for the effectiveness of RAG. This instruction fine-tuning does not always need to be done with domain-specific data.

3. Methodology

3.1. Language Models

For the experiments in this study, we used a Portuguese-adapted version of the model T5 [Raffel et al., 2020], called PTT5 [Carmo et al., 2020], and two versions of the Llama-3 8B model [AI at Meta, 2024].

PTT5 was pre-trained on the BrWac [Wagner et al., 2018], a dataset composed of millions of Internet pages in Brazilian Portuguese. This study used the base version of the model, which contains 220M parameters. This choice was due to its relatively small size, Portuguese pre-training, and encoder-decoder architecture, distinguishing it from Llama 3. Llama 3, which was developed by Meta, uses a decoder-only architecture and is available in both pre-trained and instruction-tuned versions. Despite being trained mainly on English data, Llama 3 is multilingual and was evaluated exclusively with Portuguese data in this study. The 8-billion-parameter version of Llama was chosen for its suitability to limited computational resources and its popularity as a widely used open-source model.

3.2. Fine-tuning Approach

For fine-tuning the models, we used two different techniques: full parameter fine-tuning and the QLoRa technique [Dettmers et al., 2023]. Full fine-tuning was

applied only to the PTT5 model, due to its reduced size and the limited availability of computational resources. For the Llama 3-8B models, we chose the QLoRa technique due to the large number of parameters in these models.

3.3. RAG Setup

We adopted a basic RAG setup, in which we used the retriever and language models with no changes to their original architectures. Three retriever models were evaluated: BM25 [Robertson and Zaragoza, 2009], Dense Passage Retriever (DPR) [Karpukhin et al., 2020], and ColBERT [Khattab and Zaharia, 2020]. With DPR, the embeddings were generated by a BERT-based [Devlin et al., 2019] model known as Sentence-BERT [Reimers and Iryna, 2019]. In the case of ColBERT, we specifically used its second version - ColBERTv2 [Santhanam et al., 2021]. All IR models were used without any type of training on the data under study. The texts retrieved for each query were concatenated and inserted into the input prompt of the LLMs as support texts.

3.4. Evaluation Setup

We considered four metrics to evaluate the results of the experiments: (i) Rouge-1 and (ii) Rouge-L [Lin, 2004], which are based on word overlap; (iii) BERTScore [Zhang et al., 2019], which employs embeddings generated by a BERT model; and (iv) a specific metric developed with the use of GPT (GPT 4o mini).

GPT was used to verify the accuracy of the answers generated by the models. To achieve this, we developed the GPTScore metric, which evaluates whether the models' answers align with the content of the reference answers, even when they differ in wording, length, or style. This evaluation used the prompt shown in Figure 1. The metric was computed by tallying the number of "yes" or "no" answers provided by GPT.

Você está avaliando a saída de um modelo de linguagem de perguntas e respostas. O modelo recebe uma pergunta e um contexto com informações que ajudam a responder à pergunta. O modelo responde com base no contexto oferecido. Abaixo estão apresentados o contexto, a pergunta, a resposta dada pelo modelo e a resposta correta de referência. Você deve avaliar se a resposta dada pelo modelo contém informações equivalentes às da resposta de referência. As respostas não precisam ser idênticas e podem apresentar a mesma informação de formas diferentes; você deve avaliar somente se as informações apresentadas na resposta são equivalentes e não a forma de apresentação das informações. Responda apenas com "sim" ou "não".

Contexto: {contexto}
 Pergunta: {pergunta}
 Resposta do modelo: {resposta_modelo}
 Resposta referência: {resposta_referência}

Figure 1. Prompt used to obtain the GPTScore.

4. Experiments and Results

This section outlines the experimental setup and presents the results obtained using the RAG and fine-tuning strategies.

4.1. Experimental Setup

We utilized two datasets in Portuguese. The first one, Pirá 2.0 [Paschoal et al., 2021; Pirozelli et al., 2024], focuses on topics related to the Brazilian coast, oceans and climate change. This dataset contains questions, answers, and support texts derived from the abstracts of scientific papers and specialized reports on the aforementioned topics, all of which have a version in Portuguese. Comprising 2258 samples, this dataset was split as follows: 80% for training, 10% for validation, and 10% for testing. We selected this dataset because it includes texts in Portuguese and focuses on a specific domain. The second dataset is a Portuguese translation of the Databricks-Dolly dataset [Conover et al., 2023], which consists of pairs of instructions and answers across various task categories generated by Databricks employees. For this study, we only kept the records classified as "closed QA" task, which involves questions and answers based on excerpts from Wikipedia. This dataset contains 1766 records, distributed as follows: 70% for the training, 15% for validation, and 15% for test. It was included in this study because it contains questions from a broader domain that still require supporting texts for answer formulation.

The experiments conducted in this study aimed to explore different strategies for adapting LLMs to new domains, particularly focusing on fine-tuning and RAG-based approaches. We investigated various settings regarding the use of supporting texts, both in the fine-tuning process and during the validation stage. For the fine-tuning experiments, we evaluated two approaches: one that includes supporting texts and question-answer pairs in the input prompt, termed "RAG FT"; and another that utilizes only question-answer pairs, referred to as "QA FT". For validation, the scenarios in which support texts were included in the input prompt were called "RAG". For settings in which only the question was used as input, the following prompt was utilized: "Responda à pergunta de forma sucinta.\n\nPergunta: {question}". In cases where support texts were also included, the prompt was modified to: "Responda à pergunta de forma sucinta e com base no contexto dado. Contexto: {context}\n\n Pergunta: {question}".

In our experiments, we employed a "greedy" decoding strategy, selecting the token with the highest probability during generation. We established a maximum output limit of 100 tokens, while the input limit was set at 1024 tokens to accommodate most of the supporting texts without truncation. Fine-tuning of the Llama models employed the QLoRA method with 4-bit quantization over 10 epochs, utilizing a batch size and gradient accumulation of 4 to optimize hardware capacity. The model generated answers at 16-bit precision. Meanwhile, the PTT5 model underwent full fine-tuning for 60 epochs, using a batch size of 8 and gradient accumulation of 4. All training was conducted using Hugging Face libraries on an NVIDIA GeForce RTX 4090 GPU.

To select the best retriever for the RAG experiments, we used GPT 4o mini to answer the questions in each dataset based on the context provided by each retriever. For each question, the texts retrieved by each model were concatenated and added to the prompt used by GPT to generate the answer. For the Pirá dataset, we used the four most relevant passages identified by each retriever, while for the Dolly dataset, we used the three most relevant passages. ColBERT outperformed all other models across the evaluated metrics and was, therefore, chosen for this study. Table 1 summarizes these results and includes a hypothetically ideal retriever, simulated by using context texts that are always correct for each question.

Table 1. Evaluation of the retriever methods (see text).

GPT4o-mini		Rouge-1 (F1)	Rouge-L (F1)	BertScore (F1)	GPTScore (Acc)
BM25	Pirá	0.2660	0.2344	0.7404	0.5947
	Dolly	0.3273	0.2807	0.7539	0.5819
DPR	Pirá	0.2142	0.1823	0.7146	0.4273
	Dolly	0.3373	0.2924	0.7594	0.6328
ColBERT	Pirá	0.2723	0.2457	0.7476	0.6872
	Dolly	0.3410	0.2959	0.7612	0.6525
Ideal	Pirá	0.3185	0.2831	0.7670	0.9295
	Dolly	0.4145	0.3658	0.7947	0.9068

4.2. Models without fine-tuning

The experiments with models without fine-tuning, considering only the question in the input prompt, may reveal their level of prior knowledge about the datasets' domain. The results for this setting, referred to as "No FT, No RAG" in Table 2, indicate that these models have low prior knowledge of the Pirá dataset's domain and moderate knowledge of the Dolly dataset. In this analysis, results for PTT5 models are not reported, as we were unable to obtain satisfactory answers from this model without fine-tuning.

In the case of RAG experiments with models without fine-tuning, referred to as "No FT, RAG", we observed an increase in GPTScore and a more modest rise in Rouge metrics. This behavior is expected, as Rouge metrics, which assess term matching, are more influenced by the style of the answers – particularly their length and vocabulary. Since the models were not fine-tuned to the datasets of interest, the generated answers may not align with the answer patterns from the dataset. It was observed that the pre-trained model often answered the questions and then continued generating question-answer pairs indefinitely until it reached the maximum number of output tokens. The Llama Instruct model performed significantly better on both the GPTScore and BERTScore due to its prior fine-tuning, which enhanced its reading comprehension abilities. This suggests that

models with advanced comprehension skills, even if trained in domains different from those being tested, can substantially benefit from the use of supporting contexts to leverage their performance. We can also observe that models without prior domain knowledge and that did not explore RAG were the worst performers.

4.3. Models fine-tuned solely with question-answer pairs

This setting aims to evaluate whether the models can internalize knowledge about the domains through the fine-tuning process, specifically based on questions and answers from the datasets, referred to as "QA FT, No RAG" in Table 2. The results show that for both datasets, fine-tuning does not provide a significant improvement when the model is tested without RAG. However, when the model includes RAG, referred to as "QA FT, RAG" in Table 2, we observed a meaningful gain in some evaluation scenarios. This suggests that the fine-tuning process helps the model learn the style of the answers - the length of the answers becomes more similar to that observed in the dataset - but does not necessarily enable it to retain domain knowledge. It is worth noting that the limited amount of training data may hinder the model's ability to learn effectively through the fine-tuning process.

4.4. Models fine-tuned with question, answer and context

In this scenario, the models were fine-tuned with the addition of contexts in the training prompts. In the validation setting without RAG, referred to as "RAG FT, No RAG", all models performed poorly. This result is expected, as the primary purpose of fine-tuning with added contexts is to train the model to generate answers based on the context itself. Since this setting does not include the support texts in the input prompts, fine-tuning did not appear to achieve the desired outcome.

In the setting that includes RAG, referred to as "RAG FT, RAG", the models achieved the best results across all metrics for the two datasets analyzed. It is worth noting that for the Llama Instruct model, which was already fine-tuned for the reading comprehension task, all settings that utilized RAG performed well according to GTPScore. However, for Rouge metrics, the models with fine-tuning on domain-specific data showed superior performance. This experiment suggests that even if the model is capable of extracting answers from the context, fine-tuning on problem-specific data may be beneficial for generating answers in a format more closely aligned with that found in the dataset. We also observed that fine-tuning the pre-trained Llama model allowed it to achieve results comparable to those of the Llama Instruct model, despite the latter being previously fine-tuned with a significantly larger amount of data. This result indicates that fine-tuning with context texts, even when performed with a reduced dataset, can enhance the model's ability to extract relevant information from context. In this setting, we also observed a significant improvement in the PTT5 results, which were clearly surpassed by those obtained with the Llama models, likely to their much larger number of parameters.

Table 2. Experimental results (see text).

Llama 3 PT (Pirá)	Rouge-1 (F1)	Rouge-L (F1)	BertScore (F1)	GPTScore (Acc)	Llama 3 PT (Dolly)	Rouge-1 (F1)	Rouge-L (F1)	BertScore (F1)	GPTScore (Acc)
No FT, No RAG	0.1091	0.0967	0.6305	0.0264	No FT, No RAG	0.1638	0.1420	0.6233	0.1412
No FT, RAG	0.1772	0.1525	0.6795	0.3624	No FT, RAG	0.2375	0.2029	0.6799	0.3505
QA FT, No RAG	0.1784	0.1628	0.7245	0.1101	QA FT, No RAG	0.2807	0.2464	0.7477	0.1554
QA FT, RAG	0.3322	0.3105	0.7802	0.4978	QA FT, RAG	0.3579	0.3238	0.7713	0.4718
RAG FT, No RAG	0.1420	0.1301	0.7101	0.0352	RAG FT, No RAG	0.2489	0.2204	0.7386	0.1554
RAG FT, RAG	0.3711	0.3521	0.7905	0.5903	RAG FT, RAG	0.3815	0.3489	0.7805	0.5339
Llama 3 Instruct (Pirá)	Llama 3 Instruct (Dolly)								
No FT, No RAG	0.1603	0.1358	0.6940	0.0529	No FT, No RAG	0.2498	0.2159	0.7305	0.2599
No FT, RAG	0.2824	0.2541	0.7458	0.6035	No FT, RAG	0.3166	0.2850	0.7498	0.5678
QA FT, No RAG	0.1830	0.1661	0.7236	0.0969	QA FT, No RAG	0.2798	0.2433	0.7502	0.1695
QA FT, RAG	0.3538	0.3307	0.7831	0.5859	QA FT, RAG	0.3698	0.3385	0.7756	0.5424
RAG FT, No RAG	0.1587	0.1438	0.7136	0.0793	RAG FT, No RAG	0.2600	0.2268	0.7438	0.1667
RAG FT, RAG	0.3699	0.3520	0.7939	0.6035	RAG FT, RAG	0.3967	0.3593	0.7867	0.5597
PTT5 (Pirá)	PTT5 (Dolly)								
QA FT, No RAG	0.1859	0.1708	0.7230	0.0925	QA FT, No RAG	0.2228	0.1971	0.7212	0.0169
QA FT, RAG	0.1379	0.1188	0.6666	0.0396	QA FT, RAG	0.1892	0.1524	0.6647	0.1158
RAG FT, No RAG	0.1282	0.1164	0.6929	0.0132	RAG FT, No RAG	0.2028	0.1837	0.7110	0.0198
RAG FT, RAG	0.3028	0.2859	0.7642	0.3744	RAG FT, RAG	0.3134	0.2713	0.7478	0.2429

5. Conclusion

This work analyzed various methods for adapting LLMs to specific domains in QA tasks, including fine-tuning the model and integrating external data through RAG. The experiments demonstrated that incorporating external data generally improves the models' performance, regardless of whether fine-tuning is applied. The results also showed that fine-tuning, even when conducted with a reduced dataset, can enhance the models' performance. Additionally, we observed that while the best results were achieved by models specifically tuned to domain data, a model with previously fine-tuned instructions produced similar outcomes, with the clear advantage of not requiring any additional fine-tuning.

The experiments presented here were conducted using a basic RAG architecture, without any additional training of the retrievers on the datasets of interest. Future work could explore the same settings with adjustments to the retrievers as well.

6. Acknowledgments

To CNPq, FAPERJ, and CAPES. This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) – Finance Code 001.

7. References

- Brown, Tom B. "Language models are few-shot learners." *In Proceedings of the 34th International Conference on Neural Information Processing Systems (NIPS '20)* (2020).
- Achiam, Josh, OpenAI et al. "GPT-4 technical report." *arXiv preprint arXiv:2303.08774* (2023).
- Kandpal, Nikhil, et al. "Large language models struggle to learn long-tail knowledge." *International Conference on Machine Learning*. PMLR, (2023).
- Kasai, Jungo, et al. "REALTIME QA: What's the Answer Right Now?" *Advances in Neural Information Processing Systems* 36 (2024).
- Zhang, Tianjun, et al. "RAFT: Adapting Language Model to Domain Specific RAG." *arXiv preprint arXiv:2403.10131*, (2024).
- Guo, Kunpeng, et al. "Fine-tuning Strategies for Domain Specific Question Answering under Low Annotation Budget Constraints." *IEEE 35th International Conference on Tools with Artificial Intelligence (ICTAI)*. IEEE, (2023).
- Lewis, Patrick, et al. "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks." *Advances in Neural Information Processing Systems* 33: 9459-9474, (2020).
- Dettmers, Tim et al. "QLoRA: Efficient Finetuning of Quantized LLMs." *ArXiv abs/2305.14314*, (2023).
- Balaguer, Angels, et al. "RAG vs fine-tuning: Pipelines, tradeoffs, and a case study on agriculture." *arXiv e-prints* (2024): arXiv-2401.
- Lála, Jakub, et al. "PaperQA: Retrieval-augmented generative agent for scientific research." *arXiv preprint arXiv:2312.07559*, (2023).
- Zakka, Cyril, et al. "Almanac—retrieval-augmented language models for clinical medicine." *NEJM AI* 1.2 (2024): AIoa2300068.
- Borgeaud, Sebastian, et al. "Improving language models by retrieving from trillions of tokens." *International conference on machine learning*. PMLR, (2022).
- Shuster, Kurt, et al. "Retrieval augmentation reduces hallucination in conversation." *arXiv preprint arXiv:2104.07567*, (2021).
- Vu, Tu, et al. "FreshLLMs: Refreshing large language models with search engine augmentation." *arXiv preprint arXiv:2310.03214*, (2023).
- Izacard, Gautier, et al. "Atlas: Few-shot learning with retrieval augmented language models." *Journal of Machine Learning Research* 24.251 (2023): 1-43.
- Sparck Jones, Karen. "A statistical interpretation of term specificity and its application in retrieval." *Journal of documentation* 28.1 (1972): 11-21.

- Robertson, Stephen E. and Hugo Zaragoza. "The Probabilistic Relevance Framework: BM25 and Beyond." *Found. Trends Inf. Retr.* 3 (2009): 333-389.
- Karpukhin, Vladimir, et al. "Dense passage retrieval for open-domain question answering." *arXiv preprint arXiv:2004.04906* (2020).
- Khattab, Omar, and Matei Zaharia. "ColBERT: Efficient and effective passage search via contextualized late interaction over BERT." *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval* (2020).
- Xu, Lingling, et al. "Parameter-efficient fine-tuning methods for pretrained language models: A critical review and assessment." *arXiv preprint arXiv:2312.12148* (2023).
- Hu, Edward J., et al. "LoRA: Low-rank adaptation of large language models." *arXiv preprint arXiv:2106.09685* (2021).
- Li, Yixiao, et al. "LoftQ: Lora-fine-tuning-aware quantization for large language models." *arXiv preprint arXiv:2310.08659* (2023).
- Ovadia, Oded, et al. "Fine-tuning or retrieval? comparing knowledge injection in LLMs." *arXiv preprint arXiv:2312.05934* (2023).
- Raffel, Colin, et al. "Exploring the limits of transfer learning with a unified text-to-text transformer." *Journal of machine learning research* 21.140 (2020): 1-67.
- Carmo, Diedre, et al. "PTT5: Pretraining and validating the T5 model on Brazilian Portuguese data." *arXiv preprint arXiv:2008.09144* (2020).
- "Introducing Meta Llama 3: The most capable openly available LLM to date." AI at Meta. (2024). <https://ai.meta.com/blog/meta-llama-3/>
- Wagner Filho, Jorge A., et al. "The brWaC corpus: A new open resource for Brazilian Portuguese." *Proceedings of the eleventh international conference on language resources and evaluation LREC* (2018).
- Devlin, Jacob et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." *North American Chapter of the Association for Computational Linguistics* (2019).
- Reimers, Nils and Iryna Gurevych. "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks." *Conference on Empirical Methods in Natural Language Processing* (2019).
- Santhanam, Keshav et al. "ColBERTv2: Effective and Efficient Retrieval via Lightweight Late Interaction." *North American Chapter of the Association for Computational Linguistics* (2021).

- Lin, Chin-Yew. "ROUGE: A Package for Automatic Evaluation of Summaries." *Annual Meeting of the Association for Computational Linguistics* (2004).
- Zhang, Tianyi, et al. "BERTScore: Evaluating text generation with BERT." *arXiv preprint arXiv:1904.09675* (2019).
- Paschoal, André FA, et al. "Pirá: A bilingual portuguese-english dataset for question-answering about the ocean." *Proceedings of the 30th ACM International Conference on Information & Knowledge Management* (2021).
- Pirozelli, Paulo, et al. "Benchmarks for Pirá 2.0, a Reading Comprehension Dataset about the Ocean, the Brazilian Coast, and Climate Change." *Data Intelligence* 6.1 (2024): 29-63.
- Conover, Mike, et al. "Free Dolly: Introducing the world's first truly open instruction-tuned LLM." *Company Blog of Databricks* (2023). <https://www.databricks.com/blog/2023/04/12/dolly-first-open-commercially-viable-instruction-tuned-llm>

A Change in Perspective: The Trade-Off Between Perspective API and Custom Models in Classifying Hate Speech in Portuguese

Arthur Buzelin¹, Yan Aquino¹, Pedro Bento¹, Samira Malaquias¹,
Wagner Meira Jr¹, Gisele L. Pappa¹

¹Departamento de Ciência da Computação – Universidade Federal de Minas Gerais
Belo Horizonte – MG – Brazil

{arthurbuzelin, yanaquino, pedro.bento, samiramalaquias}@dcc.ufmg.br
{meira, glpappa}@dcc.ufmg.br

Abstract. *This paper examines the performance of the Perspective API, developed by Jigsaw, in detecting hate speech in Portuguese. Although the Perspective API supports multiple languages, its performance metrics are often aggregated, obscuring specific details. Our study reveals that the API's AUC-ROC score for Portuguese is significantly lower than for English (0.744 vs. 0.942). To address this, we developed a BERT classifier model trained on a Portuguese Twitter hate speech dataset. Our model, with just 100 messages in its training set, outperformed the Perspective API. These findings highlight the need for more granular performance metrics and suggest that custom models may offer better solutions for specific languages.*

1. Introduction

Perspective API is a tool designed to identify and mitigate toxic language online [Lees et al. 2022]. Using advanced machine learning and Natural Language Processing (NLP) models, Perspective API analyzes textual content to detect various forms of harmful speech, including threats, insults, and hate speech. It is considered state-of-the-art for detecting toxicity, and used by multiple platforms, such as Reddit, The New York Times, The Wall Street Journal, and EL PAÍS.

Despite its widespread adoption and claimed multilingual support, including Portuguese, the actual performance of the Perspective API in different languages remains unclear. The official documentation and associated research papers often report performance metrics by aggregating data from multiple languages, within a multilingual dataset. This aggregation conceal the individual performance metrics for Portuguese, making it difficult to evaluate the API's effectiveness in this specific language. The lack of transparency in language-specific performance metrics raises concerns about the API's reliability when applied to non-English texts.

The widespread acceptance of Perspective API as a leading tool for hate speech detection combined with its claimed support for Portuguese, suggests that professionals may readily adopt it for work in Portuguese-speaking contexts. However, if the API's performance in Portuguese is not on par with its performance in English or the aggregated results, this could lead to inaccurate analyses and conclusions, particularly in fields like

computational social sciences, where precise language detection is critical. The potential for misleading results is especially concerning when the tool’s reliability in Portuguese is taken for granted based on its performance in other languages.

This paper addresses this gap by evaluating the performance of the Perspective API in detecting hate speech specifically in Portuguese. It also assesses the feasibility of developing a custom hate speech detection tool tailored for Portuguese. To guide our investigation, we formulated the following research questions:

RQ1: How well does the Perspective API perform when detecting hate speech in Portuguese?

RQ2: For Portuguese, is it more effective and efficient to use a custom-made tool rather than relying on existing solutions like the Perspective API? If so, how much effort would it take to build it?

To address these questions, we evaluated the Perspective API’s metrics using a Portuguese Twitter hate speech dataset. We then compared it to metrics obtained in a similar English dataset regarding classification, date of collection, and content. Our findings revealed that the Perspective API’s performance in Portuguese was significantly worse than in English. Based on this insight, we developed our version of a BERT classifier to detect hate speech in Portuguese. Remarkably, with just 100 messages, the BERT model outperformed the Perspective API in detecting hate speech in Portuguese. In contrast, the BERT model trained with the English dataset did not surpass the Perspective API’s performance.

2. Related Works

This section reviews studies related to hate speech detection models and language-specific performance comparisons. opportunities

2.1. Model Comparisons

Multilingual transformer models, such as BERT and its variants, have gained significant attention in hate speech detection across various languages. For instance, [Roy et al. 2021] demonstrated the superiority of fine-tuned transformer models in handling multilingual data, showcasing their effectiveness compared to more generalized approaches like the Perspective API. This highlights the potential of specialized models to outperform broader, one-size-fits-all solutions.

Another noteworthy contribution is by [Kennedy et al. 2020], who introduced a hybrid approach that combines faceted Rasch measurement with multitasking deep learning. This methodology enhances both the interpretability and precision of hate speech detection by integrating traditional psychometric techniques with advanced deep learning models. Compared to the Perspective API, which relies on more generalized algorithms, Kennedy et al.’s approach offers a more nuanced understanding of linguistic variations and the intensity of hate speech.

2.2. Language-Specific Comparisons

In Perspective’s introductory paper [Lees et al. 2022], developers reported AUC-ROC scores of 0.98 for English, 0.91 for Russian, and 0.87 for a group of ten other languages.

These results highlight a disparity in the API’s effectiveness across languages, raising concerns about its applicability in non-English contexts.

Further studies have confirmed these concerns were relevant. For instance, [Nogara et al. 2024] analyzed the use of the Perspective API in German and found that the API tends to classify German texts as significantly more toxic than their English counterparts. This finding underscores the potential biases and inaccuracies that arise when applying the API to languages other than English, highlighting the need for further investigation into its multilingual capabilities.

The seminal study of the use of the Perspective API’s in Portuguese was conducted by [Kobellarz and Silva 2022]. They compared identical texts in Portuguese and English using the API and concluded that it performs better when analyzing texts in their original language. This suggests that the Perspective API may be less effective in detecting nuances in translated or non-native language content.

Building upon this study, [Lima et al. 2024] developed a manually labeled dataset of toxic messages in Portuguese and evaluated the API against this dataset. Their findings revealed significant discrepancies, emphasizing the need for the API to undergo more focused training on Portuguese-language content to improve its accuracy and reliability in detecting hate speech.

Additionally, [Silva et al. 2023] proposed standardized datasets and benchmarks for sentiment analysis in English, specifically addressing the challenges of automating the development process. While their focus was on English, the methods and standards they advocate could provide valuable insights for improving the Perspective API’s performance in other languages, including Portuguese.

2.3. Research Gap

Despite the widespread use and validation of the Perspective API for hate speech detection in various languages, a significant gap remains in its performance evaluation for less commonly studied languages like Portuguese. Previous research has shown the API’s strong performance in English and other major languages, demonstrated by high AUC-ROC scores and robust metrics. However, detailed assessments for less-represented languages in its training datasets are lacking.

To address these gaps, we conducted focused evaluations of the Perspective API’s performance for individual languages. Our study highlights the advantages of developing custom models tailored to specific languages, such as Portuguese, offering more accurate and reliable hate speech detection. This emphasizes the need to consider custom solutions alongside existing multilingual models to improve the effectiveness of hate speech detection across diverse languages.

3. Methodology

In this section, we discuss the dataset selection, Perspective API evaluation, and the BERTs models fine-tuning.

3.1. Dataset

Our analysis required a Portuguese hate speech dataset and a similar English dataset, for the purpose of an unbiased comparison. Instead of manually labeling messages,

which can be costly and prone to errors, we opted to use two well-known Twitter hate speech datasets: the Hierarchically-Labeled Portuguese Hate Speech Dataset [Fortuna et al. 2019] and the Automated Hate Speech Detection and the Problem of Offensive Language dataset [Davidson et al. 2017].

Both datasets were created using the same methodology for classifying messages. This involved identifying and mining accounts likely to post hate speech-related tweets in 2017. The tweets were then classified as either containing hate speech or not, which matches the output of the Perspective API.

The original Portuguese and English datasets vary significantly in size and proportion of hate speech messages. The Portuguese dataset includes 5,934 non-toxic messages and 1,607 toxic messages, resulting in a ratio of approximately 3.7 non-toxic messages per toxic message. On the other hand, the English dataset initially consisted of 25,000 classified tweets, with 3,280 non-toxic messages and 21,720 toxic messages.

For a fair comparison of classification scores between the two datasets, we balanced their proportions by using the Portuguese dataset as the baseline, since this will be the main object of our study. By selecting a random sample of messages from the English dataset that reflected the same proportion, we leveraged a final English dataset consisting of 3,280 non-toxic messages and 886 toxic messages, with the same ratio of non-toxic to toxic messages of approximately 3.7.

3.2. Comparing Perspective API results

To compare the models of Perspective for English and Portuguese, we selected random samples of messages and analyzed them for toxicity using the Perspective API. We focused on the *Toxicity* attribute, which is widely used in literature due to its robustness and compatibility with both datasets under examination. The analysis was conducted in June 2024, and the Perspective API provided toxicity scores for each sample in both datasets.

Each message was assigned a toxicity score ranging from 0 to 1, where 0 represents a very low probability of toxicity and 1 indicates a very high probability. To ensure the most precise possible comparison, we optimized the threshold for toxicity classification by maximizing the F1 score for each dataset individually. The optimal threshold was determined to be 0.48 for the Portuguese dataset and 0.59 for the English dataset, reflecting the different calibrations needed by the two languages.

3.3. BERThs (BERT hate speech) Model

This section shows how we fine-tuned our own BERT classifier for hate speech detection, namely BERThs. BERThs was fine-tuned using both a Portuguese and an English dataset.

Initially, the goal of the model, particularly the Portuguese one, was not to achieve the highest possible accuracy, but to be easy to replicate. This will help us show whether a simple fine-tuned model may be more effective than the Perspective API in Portuguese.

For fine-tuning the BERThs-Pt, we used BERTimbau [Souza et al. 2020] as the base model, as it is pre-trained in Portuguese and better suited for our task. Given the small size of the annotated corpus, we fine-tuned and evaluated the model 30 times using different randomized non-overlapping stratified sets: training, validation, and test sets, comprising 80%, 10%, and 10% of the labeled dataset. Each split maintained the original

class distribution of approximately 21.3% toxic messages and 78.7% non-toxic messages. The same test sets were used to evaluate the Perspective API. This approach ensured robustness and prevented issues such as training on an all-toxic set of messages, which could lead to unreliable results.

To determine the minimum number of messages needed for our classifier to outperform the Perspective API, initially, only 10 messages from the training set were used for fine-tuning BERTimbau. We incrementally added 10 more messages to the training set after each iteration, until BERThs achieved a better AUC score than Perspective. The AUC metric was chosen because it was the only metric reported for Portuguese in the Perspective API paper. After that, we added 200 new messages to the training set in each subsequent iteration, until all messages were included, highlighting the highest performance our model could achieve.

The fine-tuning was performed using the PyTorch library [Paszke et al. 2019], with the AdamW optimizer [Loshchilov and Hutter 2017] and a learning rate of 5×10^{-6} . The classification thresholds were established based on the output probabilities of the model, defined as the thresholds that yielded the best mean F1-score on our validation set.

For BERThs-En, we employed the BERT uncased model [Devlin et al. 2019], which is optimized for English language processing. The fine-tuning procedure followed the same general approach used for the Portuguese variant, but with specific modifications to account for the superior performance of the Perspective API on English texts. Specifically, instead of gradually increasing the training set by 10 messages and subsequently by 200 messages per iteration, we opted to directly increase the training set by 200 messages in each iteration.

4. Results

This section presents the Perspective API prediction metrics for the English and Portuguese datasets and compares them to BERThs. The models were fine-tuned on an NVIDIA RTX 4090 GPU. As the models were trained 30 times with different data samples, the results in this section present the mean followed by the standard deviation.

4.1. Perspective Performance

Table 1 shows the performance metrics of the Perspective API in the English and Portuguese datasets. It highlights a significant disparity in the model’s effectiveness between the two languages, with the English dataset consistently achieving higher scores across all metrics. Notably, the accuracy, precision, recall, F1 score, and AUC-ROC are considerably lower for the Portuguese dataset, suggesting that the model’s capability to accurately classify toxic content is compromised in Portuguese.

Note that the F1 score – which serves as a balanced measure of a model’s precision and recall in classification tasks – is almost 35 percentage points lower in Portuguese. On top of that, typically, there is a trade-off between the precision and recall metrics; adjusting the threshold to improve one often causes the deterioration of the other. However, in this case, precision and recall are significantly lower for the Portuguese dataset, indicating an overall performance issue. Low precision usually implies in a high number of false positives, while low recall indicates many false negatives.

Table 1. Metrics for the Perspective API model in English and Portuguese.

Metric	Perspective API(En)	Perspective API(Pt)	Difference
Accuracy	0.901	0.779	0.122
Precision	0.813	0.477	0.336
Recall	0.744	0.404	0.340
F1 Score	0.777	0.438	0.339
AUC-ROC	0.942	0.743	0.199

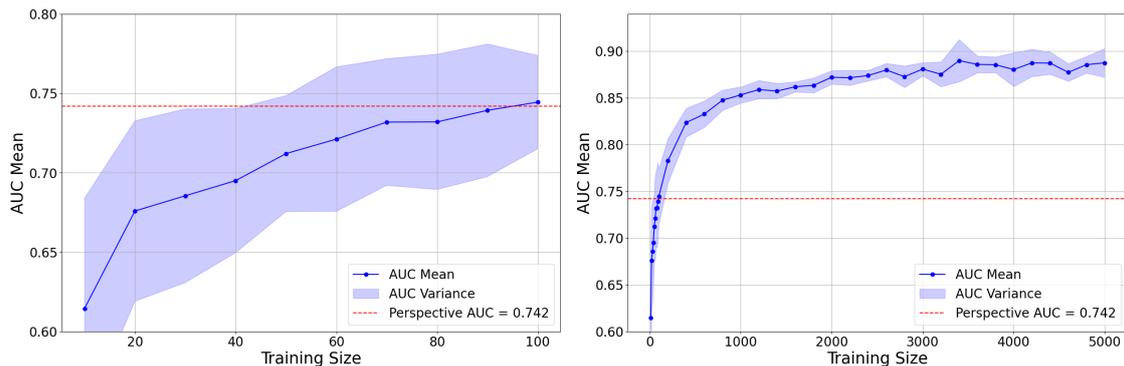


Figure 1. Graph displaying the mean AUC across varying training sizes of the BERTs-Pt model, with the left panel covering up to 100 Twitter posts to assess early performance, and the right panel extending to 5000 posts to evaluate the model’s full training potential.

The most concerning results are in the AUC-ROC score, which measures the classification abilities of the Perspective API in its official paper. The Portuguese dataset scores 20 percentage points lower than the English dataset in AUC-ROC. This is both surprising and alarming, given that the multilingual Perspective API is reported to have an AUC-ROC of 0.877, only slightly lower than the English counterpart in the official documentation.

These findings suggest that, while the Perspective API claims to support multiple languages, its performance in Portuguese is substantially lower than in English. This underscores the importance of evaluating multilingual models on a per-language basis to ensure their effectiveness and reliability across different linguistic contexts.

4.2. Evaluating BERTs

BERTs-Pt was evaluated by incrementally increasing the training set by 10 messages at a time. Figure 1 illustrates the model’s performance as the number of training messages increased. The red line represents the average performance of the Perspective API on the test set. Observe that our model surpassed the Perspective API in AUC-ROC score with only 100 training messages.

On the other hand, the BERTs-En dataset showed a different result. Even with an extensive training data, leaving aside a small portion for testing and validation, the fine-tuned BERT model still performed worse than the Perspective API, achieving an average AUC-ROC of 0.934 compared to Perspective’s 0.942.

These findings suggest that while the Perspective API is an excellent tool, it is

not suited for Portuguese. This means creating a custom model can easily surpass the Perspective API’s performance with a relatively small amount of labeled data. Therefore, a classifier tailored to the specific linguistic and contextual nuances of Portuguese is better suited for the detection of hate speech.

4.3. Qualitative Analysis

Table 2 shows a comparative analysis of the classification outcomes for the Perspective API and BERTs-Pt on a Portuguese hate speech dataset. The most notable observation from this analysis is that BERTs-Pt misclassified significantly fewer messages (14.3%) when compared to the Perspective API(20.9%), indicating that the BERT model is generally more accurate in discerning the nuances of the text. This superiority is particularly evident in its handling of subtle and context-dependent instances of hate speech, where the Perspective API often struggles. The analysis further reveals that while both models perform well with clear and unambiguous content, they encounter challenges with ambiguous language and contextually rich messages, cases where the BERT model shows a better overall ability to navigate these complexities.

Table 2. Comparison of misclassified and correctly classified messages from the Perspective API and BERTs-Pt. Four random examples from each quadrant are included. In quadrants where only one model missed the true class, the message labels refer to the model that made the mistake. “FP” refers to False Positive, “FN” to False Negative, “TP” to True Positive, and “TN” to True Negative.

	BERTs-Pt was correct	BERTs-Pt missed
Perspective was correct	<p>72.2% of Messages</p> <p>(TP) “Que mulher burra do cacete”</p> <p>(TP) “gorda e feia”</p> <p>(TN) “Boa semana para todos!”</p> <p>(TN) “Não vou orar, sou ateu”</p>	<p>6.9% of Messages</p> <p>(FP) “Nossa, mas feminismo necessário hoje em dia?”</p> <p>(FN) “Se vc bate nessa mulher, além de covarde, com certeza vc gosta de _”</p> <p>(FP) “quem é playboy safado fortalece no RT”</p> <p>(FP) “Isso sim é tratar gay com indiferença...”</p>
Perspective missed	<p>13.5% of Messages</p> <p>(FN) “Vai também ser lançado um manual de boas maneiras para lidar com fufas, gays e transsexuais, os chamados LGTB”</p> <p>(FN) “as pessoas não entendem que no meio dos refugiados tem inúmeros terroristas, é uma coisa tão óbvia”</p> <p>(FN) “E traveco mesmo , mó pirocão”</p> <p>(FN) “Você é cheinha, NÃO é gostosa.”</p>	<p>7.4% of Messages</p> <p>(FP) “Pra mim BBB sempre foi uma merda.”</p> <p>(FN) “feliz dia do não tenho roupa pra sair”</p> <p>(FN) “Isso é injusto!”</p> <p>(FN) “meritocracia: existe”</p>

Having established that BERTs-Pt generally outperforms the Perspective API, we conducted a quadrant-specific analysis to explore these differences further. The first quadrant represents messages that both methods classified correctly, accounting for 72.2% of the messages. This indicates their effectiveness in handling unambiguous content,

as shown in Table 2. The high success rate highlights the capability of both models to manage straightforward cases of hate speech or benign content where linguistic ambiguity is low. However, real-world scenarios often involve more nuanced language, where model differences become more evident.

The second quadrant covers the 6.9% of messages that Perspective correctly classified but BERTs-Pt misclassified. A random sample of four of these messages reveals that they are somewhat ambiguous, making it difficult to determine with certainty whether they were wrongly classified. These cases highlight the challenges of accurately categorizing nuanced and context-dependent language.

The third quadrant, which includes 13.5% of the messages that Perspective misclassified and BERTs-Pt correctly classified. It becomes apparent that Perspective struggles with more complex contexts, particularly when the hate speech is not explicit. The model has particular difficulty with slang or coded language, such as derogatory terms targeting LGBTQ+ individuals. Perspective’s limitations in understanding such indirect insults become evident here, suggesting that its generalized training may not sufficiently capture the nuances of the Portuguese language.

Finally, the fourth quadrant, comprising 7.4% of messages, involves cases where both models failed. These messages typically lack sufficient context, making accurate classification challenging. The shared difficulty in this category underscores the challenges of detecting hate speech when language is ambiguous or context is missing.

5. Conclusions

This study assessed the performance of the Perspective API in detecting hate speech in Portuguese, comparing it to English and exploring the potential of custom-trained models. The results show a significant performance gap, with the API achieving an AUC-ROC score of 94.2 in English but only 74.4 in Portuguese. This drop illustrates the limitations of using a generalized multilingual tool for specific languages.

Relying on a model that supports Portuguese yet delivers subpar results poses two main issues. First, research conducted using such a tool may produce inaccurate or misleading outcomes, undermining the validity of the study. Second, researchers from non-English-speaking regions, may feel compelled to conduct their research in English contexts to leverage the more reliable performance of tools like the Perspective API, potentially overlooking important linguistic and cultural nuances.

While the Perspective API excels in English, our study shows it may not be the best choice for Portuguese. A custom BERT model we developed using BERTimbau outperformed the API with only 100 training messages, suggesting that fine-tuning models for specific languages can yield better results in hate speech detection.

In conclusion, while the Perspective API offers robust performance for English, its efficacy in Portuguese is limited. Researchers and practitioners should consider developing custom models tailored to their specific linguistic contexts to achieve more accurate and reliable results.

Acknowledgments

This work was partially funded by CNPq, CAPES, FAPEMIG, and IAIA - INCT on AI.

References

- Davidson, T., Warmusley, D., Macy, M., and Weber, I. (2017). Automated hate speech detection and the problem of offensive language. In *Proceedings of the 11th International AAAI Conference on Web and Social Media (ICWSM)*. AAAI.
- Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding.
- Fortuna, P., Nunes, S., Soler-Company, J., and Wanner, L. (2019). A hierarchically-labeled portuguese hate speech dataset. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 94–104. Association for Computational Linguistics.
- Kennedy, C., Bacon, G., Sahn, A., and Vacano, C. (2020). Constructing interval variables via faceted rasch measurement and multitask deep learning: a hate speech application.
- Kobellarz, J. K. and Silva, T. H. (2022). Should we translate? evaluating toxicity in online comments when translating from portuguese to english. In *Anais do Simpósio Brasileiro de Sistemas Multimídia e Web (WebMedia)*, pages 95–104, Porto Alegre, Brazil. Sociedade Brasileira de Computação. In: 28th Simpósio Brasileiro de Sistemas Multimídia e Web (WebMedia), 2022, Curitiba.
- Lees, A., Tran, V. Q., Tay, Y., Sorensen, J., Gupta, J., Metzler, D., and Vasserman, L. (2022). A new generation of perspective api: Efficient multilingual character-level transformers.
- Lima, Q. L. H., Pagano, S. A., and da Silva, A. (2024). Toxic content detection in online social networks: A new dataset from brazilian reddit communities. In *16th International Conference on Computational Processing of Portuguese (PROPOR 2024)*.
- Loshchilov, I. and Hutter, F. (2017). Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.
- Nogara, G., Pierri, F., Cresci, S., Luceri, L., Törnberg, P., and Giordano, S. (2024). Toxic bias: Perspective api misreads german as more toxic.
- Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. (2019). Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.
- Roy, S. G., Narayan, U., Raha, T., Abid, Z., and Varma, V. (2021). Leveraging multilingual transformers for hate speech detection. *ArXiv*, abs/2101.03207.
- Silva, M., de Oliveira, V., and Pardo, T. (2023). A sentiment analysis benchmark for automated machine learning applications and a proof of concept in hate speech detection. In *Anais do XIV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 199–206, Porto Alegre, RS, Brasil. SBC.
- Souza, F., Nogueira, R., and Lotufo, R. (2020). Bertimbau: Pretrained bert models for brazilian portuguese. In *Proceedings of the 9th Brazilian Conference on Intelligent Systems (BRACIS)*, pages 403–417. IEEE.

No Argument Left Behind: Overlapping Chunks for Faster Processing of Arbitrarily Long Legal Texts

Israel Fama^{1*}, Bárbara Bueno^{1*}, Alexandre Alcoforado¹,
Thomas Palmeira Ferraz², Arnold Moya¹, Anna Helena Reali Costa¹

¹Escola Politécnica, Universidade de São Paulo (USP), São Paulo, SP, Brazil

²Télécom Paris, Institut Polytechnique de Paris, Palaiseau, France

{israelfama, barbarabueno, anna.reali}@usp.br

Abstract. *In a context where the Brazilian judiciary system, the largest in the world, faces a crisis due to the slow processing of millions of cases, it becomes imperative to develop efficient methods for analyzing legal texts. We introduce $uBERT$, a hybrid model that combines Transformer and Recurrent Neural Network architectures to effectively handle long legal texts. Our approach processes the full text regardless of its length while maintaining reasonable computational overhead. Our experiments demonstrate that $uBERT$ achieves superior performance compared to $BERT+LSTM$ when overlapping input is used and is significantly faster than $ULMFiT$ for processing long legal documents.*

1. Introduction

Legal NLP can be defined as the application of Natural Language Processing (NLP) techniques within the legal domain. This subfield of NLP has been experiencing rapidly growing interest from both academia and industry: [Katz et al. 2023] reports a significant increase in the volume of publications, rising from fewer than 30 papers in 2013 to nearly 120 in 2022. Brazil possesses the largest judiciary system in the world, comprising 18,000 judges distributed across 91 courts. At the time of writing, there are more than 84 million ongoing legal cases [CNJ 2024]. These numbers indicate both the need and the opportunity for innovative solutions to manage and analyze vast amounts of legal data.

We turn our focus to Legal Judgment Prediction (LJP), which involves predicting court decisions. Although predicting decisions may be a complex task, we argue it can be reduced to a Text Classification task, which has seen a marked increase in studies [Li et al. 2022], fueled by advancements in deep learning. In particular, the Transformer architecture emerged as a paradigm shift [Hasan 2022] for many NLP tasks. However, it still has limitations when handling long texts, which poses significant challenges in the legal domain, where documents are usually long and complex.

There is fruitful research being done on enhancing the input size limitation for Transformers, such as Retrieval-Augmented Language Models (RALMs) [Guu et al. 2020]. Current retrieval techniques, however, often trust embedding models which also can be sub-optimal when dealing with legal documents, where a single word in the whole document can make a difference. Also, these methods demand substantial computational resources and large document stores to achieve good performance.

*These authors contributed equally to this work.

Other methods combine input in a sequential way, often leveraging properties of Recurrent Neural Networks to process longer sequences [Wan et al. 2019], although those will also usually truncate the text if it is too long. But for documents in the legal domain, such as judicial decisions, most of the documents are usually composed of reasoning from the judge. Therefore, it is of our interest to have a method that uses the full text as input.

In this paper, we propose uBERT , a hybrid model that combines an encoder-based Transformer with a Recurrent Neural Network, capable of processing long texts. We propose an experimental setup with data from legal decisions, and compare uBERT to baselines BERT+LSTM, Big Bird and ULMFiT in the classification task. Our results show that uBERT slightly outperforms BERT+LSTM as long as overlapping input is introduced. Also, ULMFiT performs better for long texts, but is 4x slower than uBERT .

The remainder of this paper is structured as follows: Sect. 2 reviews related work on Legal NLP and long text classification; Sect. 3 outlines our proposal, including the formalization of the target task and the introduction of our model; Sect. 4 outlines the experiments we setup to assess our model in terms of performance and efficiency. Finally, we present the results and conclude with a discussion of the findings.

2. Related Work

Transformer-Based Approaches for Long Text Processing in Legal NLP: Longformer [Beltagy et al. 2020] employs a sparse attention mechanism, extending the input size limit to 4096 tokens, which is eight times the limit of BERT [Devlin et al. 2019]. [Hoang et al. 2023] applied this architecture to classify legal texts from the Indian Legal Documents Corpus – ILDC [Malik et al. 2021], but they did not process the entire text.

[Pappagari et al. 2019] introduced RoBERT, a method that splits long texts into overlapping chunks for recurrent encoding. While similar in concept to our architecture, a direct comparison is not possible due to limited details on their overlap and recurrence strategies. Moreover, RoBERT was evaluated on shorter texts compared to our dataset.

The overlapping algorithm in our approach, uBERT , is a specific case of SlidingBERT’s method [Zhang et al. 2023], with the stride set to half the overlap. Unlike SlidingBERT, where tokens can appear in multiple chunks, we limit overlaps to two chunks to reduce computational overhead while preserving context continuity. This choice is driven by efficiency, not language differences.

[Menezes-Neto and Clementino 2022] introduced BrCAD-5¹, a dataset designed for Legal Judgment Prediction (LJP), and evaluated three architectures for this task: ULMFiT [Howard and Ruder 2018], BigBird [Zaheer et al. 2020], and BERT+LSTM. ULMFiT, a transfer learning model that fine-tunes a pre-trained language model for downstream NLP tasks, was the only architecture capable of processing the entire text as input. BigBird, a sparse-attention model, addresses the 512-token limit by focusing on subsets of tokens, thereby reducing computational complexity, and was configured to handle texts up to 7,680 tokens. For BERT+LSTM, documents were split into 512-token chunks, with truncation applied to middle chunks if a document required more than 15. While simi-

¹This dataset consists of decisions issued by the Brazilian Federal Small Claims Court (FSCC). These decisions can be appealed to the Appellate Panel (AP), which re-examines the case and either reverses or affirms the initial ruling. Each data point in BrCAD-5 represents the text of a decision issued by the FSCC. The task proposed by the authors is to predict whether the AP will reverse or affirm the initial ruling based on the decision text.

lar in approach, uBERT differs from BERT+LSTM in that it uses a chunk overlapping strategy and imposes no limit on the number of chunks, ensuring the entire text is utilized without truncation.

Critiques and Limitations in Legal NLP Research: The legal industry has been slow to adopt NLP advancements, relying heavily on manual work by lawyers. [Mahari et al. 2023] identify a key issue: Legal NLP research often fails to align with the practical needs of legal practitioners. [Medvedeva and McBride 2023] further highlight a significant gap in Legal Judgment Prediction (LJP) research, criticizing the use of poorly designed datasets that rely on biased case facts extracted from judgments. This approach leads to models with overly optimistic performance that offer limited practical value to legal practitioners.

This work aims to bridge the gap between research and practice in the field of Legal NLP. We propose an architecture capable of processing virtually infinite-length legal texts and evaluate it on the BrCAD-5 dataset, which [Medvedeva and McBride 2023] regard as a well-designed benchmark.

3. Proposal

Text classification can be formalized as follows. Given a document d that represents a judicial decision, the goal is to make a prediction $y \in \{0, 1\}$, by learning a binary classifier f such that $f(d) = y$. The positive class $y = 1$ represents a decision that will be reversed by an Appellate Panel (AP). Since legal documents are often long, when using Transformer-based models, conventional approaches usually truncate text from d , which is sub-optimal for the task [Pappagari et al. 2019]. This can hinder performance on the Legal Judgment Prediction task, since some relevant part of the text may be cut off.

Believing that the text as a whole is more useful when learning a classifier, we propose unlimited BERT, or uBERT , an efficient architecture that combines an encoder-based Transformer with a Recurrent Neural Network, utilizing an overlapping algorithm during both training and inference to handle an unlimited number of input tokens. This approach is similar to the BERT+LSTM model used by [Menezes-Neto and Clementino 2022], but introduces modifications to maintain local context (through overlapping chunks) and accommodate documents of virtually any size. Although the quadratic memory complexity of the self-attention mechanism presents a challenge for scaling input indefinitely, we leverage the RNN’s capacity to process long sequences, enabling it to take chunk embeddings and output a comprehensive document embedding. Several studies, such as [Hoang et al. 2023], have explored the combination of attention mechanisms and recurrence. Our model builds on this concept but applies overlapping during both training and inference, and does not limit the number of chunks processed by the encoder.

Figure 1 depicts the uBERT architecture. It shows key aspects to understand how our model works.

Let E be an encoder-based Transformer, with dim being the dimensionality of the output vector of the final layer, and R be a Recurrent Neural Network. Let max_{tok} be the maximum number of tokens E can process as input. Let max_c be the maximum number of chunks of max_{tok} tokens that E can process in parallel with a single run. We split document d into n chunks of size max_{tok} tokens, starting in the first token. For each

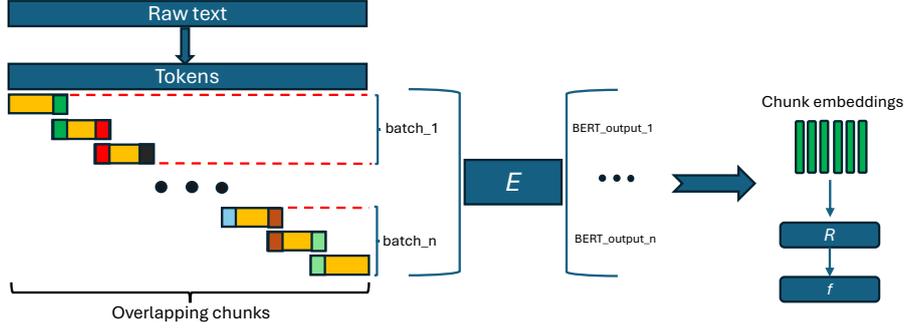


Figure 1: uBERT architecture.

run, we extract the hidden states from the last four layers of E , and concatenate them to form the representation of each chunk. This is based on the idea that different layers capture different linguistic features [Tenney et al. 2019]. Specifically, BERT+LSTM, the baseline most similar to our proposed architecture, extracts the hidden states from the last four layers. While other layers could be used for extraction, we retained this approach for consistency in model comparison. In each single run, we process $[1, max_c]$ chunks in parallel, generating $[1, max_c]$ vectors of embeddings, each with dimensionality $4 \times dim$. We iteratively process chunks from d until an embedding vector is generated for each chunk and thus preserving the entire text content of d .

Then, we concatenate the embedding vectors maintaining the order of the respective chunks, generating a tensor of dimensionality $(n, 4 \times dim)$. We process this tensor with the RNN sequentially, capturing the dependencies between them and generating a contextually enhanced representation for each chunk.

Splitting text by token count can disrupt its flow, so we use token overlap between chunks during both training and inference to maintain continuity. This technique, similar to that used by [Hoang et al. 2023] but applied more broadly, helps preserve the text’s natural structure.

Our token overlap algorithm can be formalized as follows. Consider the judicial decision d as the tokenized sequence $S = \{t_1, \dots, t_k\}$, where k is the number of tokens in d . We define the overlap size, z , as the number of tokens each chunk shares with its adjacent neighbors. Thus, any chunk shares $\lfloor \frac{z}{2} \rfloor$ tokens with the previous chunk and $\lfloor \frac{z}{2} \rfloor$ with the subsequent one. The first and last chunks, having only one adjacent chunk, share $\lfloor \frac{z}{2} \rfloor$ tokens with their respective neighbors.

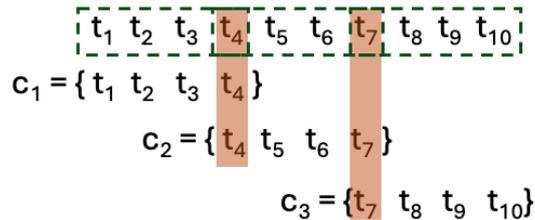


Figure 2: Overlapping chunks example.

Figure 2 provides a simple example for clarification. In this example, the chunk

size is 4, and $z = 2$. As shown, chunk $c_2 = \{t_4, t_5, t_6, t_7\}$ shares token t_4 with chunk c_1 and token t_7 with chunk c_3 . Note that the first and last chunks share only one token with the neighboring chunk.

4. Experiments

In this section, we design experiments to assess our proposed model, uBERT , and validate its effectiveness on the legal domain. We split our experiments into 3, one for each of the following research questions:

RQ1: Would an encoder-based model benefit from using the entire text in terms of performance improvement?

We examine the impact of processing the whole documents using multiple encoder passes. We first tested if simply increasing text chunks to process the whole text without using overlap (uBERT_0) improves performance over BERT+LSTM, which processes only partial text in a single pass. Then, we investigated the effect of introducing overlaps (0 to 510 tokens²) between chunks to observe if the added local context enhances predictions.

RQ2: If performance improves, does it come with reasonable computational overhead?

We compare the inference time of our architecture against all baseline models to determine if it offers a performance gain and to assess the associated computational overhead.

RQ3: Is our architecture better for processing longer texts?

We explore the relationship between document length and model performance. We tested the models on the full test set as well as on its subsets, the 10% and the 1% longest texts. This experiment involved statistical analysis to determine whether longer texts lead to better or worse predictions.

Data: We used the BrCAD-5 dataset³ in our experiments. The task is a binary classification, with Class 1 indicating that the AP reverses the previous decision, and Class 0 indicating it affirms. The dataset is imbalanced, with 22% of the data points belonging to Class 1. Although this imbalance ratio is consistent across all dataset splits, it varies significantly with text length.

Models: In this work, our model (uBERT) uses BERT as the encoder and LSTM as the RNN, with max_{tok} set to 512 tokens and up to 15 chunks (max_c) processed in parallel. Our training procedure follows the approach of [Menezes-Neto and Clementino 2022], where we fine-tune the last layer of BERT and the LSTM. The fine-tuning is conducted for 1 epoch utilizing the One Cycle learning rate scheduler. Our inference procedure mirrors the training process.

Baselines: our baseline models are ULMFiT (forward, backward and bidirectional)⁴, Big Bird and BERT+LSTM. Notably, only ULMFiT and uBERT process the full text.

²The typical input size for BERT models is 512 tokens. Our overlap algorithm first slices the text and distributes the tokens. Only after this process are the special tokens [CLS] and [SEP] added, resulting in the well-known 512-token limit.

³This dataset is divided in training, validation, and test sets: the training set includes 380,673 documents, while the validation and test sets contain 76,342 and 76,299 entries, respectively.

⁴ULMFiT incorporates a forward language model (predicting the next token), a backward language model (predicting the previous token), and a bidirectional model that combines the two, allowing it to capture contextual information in both directions.

Computational Infrastructure and Resources: the experiments were conducted using Google’s Colab infrastructure, specifically an NVIDIA A100 GPU with 40 GB of RAM.

Evaluation Metrics: We evaluate all models using the Macro F1 score and Matthews Correlation Coefficient (MCC). The Macro F1 score is a well-established metric across NLP fields, representing the harmonic mean of precision and recall, while MCC, though less common, is frequently used in the Legal Judgment Prediction (LJP) subfield, as noted by [Cui et al. 2022]. MCC measures the correlation between predicted and actual classifications by accounting for true positives, true negatives, false positives, and false negatives, making it suitable for imbalanced classes⁵. Additionally, MCC is the metric used by [Menezes-Neto and Clementino 2022], making it necessary for us to use it as well for model comparison. To compare different baselines and configurations of our uBERT model, we employed bootstrap resampling to obtain 95% confidence intervals, followed by Wilcoxon-Holm post-hoc analysis to assess statistical significance with $\alpha = 5\%$, following similar approaches [Demšar 2006, Zhu et al. 2020, Ferraz et al. 2021].

5. Results

Table 1 presents the results for all model configurations on the full test dataset, as well as the 10% and 1% longest texts. The baseline models were not run on the full test set in this study due to computational resource limitations. The results reported here are reproduced from [Menezes-Neto and Clementino 2022], which is why Table 1 does not include inference times for the full test set. Figure 3 displays the macro F1-scores across varying text lengths, while Figure 4 ranks the models using the MCC metric. Although MCC is effective for within-dataset comparisons, it is less suitable across datasets with differing class imbalance; hence, we rely on the macro F1-score for cross-dataset comparisons.

Table 1: uBERT performance across various overlap sizes compared with baselines.

Dataset:	Full Test Set (76,299 documents) Imbalance Ratio = 0.28		10% Set (7,632 documents) Imbalance Ratio = 0.32			1% Set (763 documents) Imbalance Ratio = 0.54		
	Macro-F1↑	MCC↑	Macro-F1↑	MCC↑	Inf.Time↓	Macro-F1↑	MCC↑	Inf.Time↓
<i>Baselines</i>								
ULMFiT - fwd	65.1 %	0.32	64.9 %	0.32	1h 18min	72.3 %	0.47	11min:22s
ULMFiT - bwd	65.7 %	0.35	63.4 %	0.35	1h 18min	59.9 %	0.33	14min:44s
ULMFiT - bidir	66.9 %	0.37	64.8 %	0.34	1h 18min	69.3 %	0.43	14min:44s
Big Bird	52.0 %	0.27	44.0 %	0.23	22min	30.0 %	0.08	2min:58s
BERT+LSTM	64.1 %	0.33	63.2 %	0.31	12min	64.0 %	0.36	1min:29s
<i>Ours</i>								
uBERT_0	63.9 %	0.33	62.6 %	0.31	13min	61.3 %	0.34	1min:51s
uBERT_150	63.3 %	0.32	62.2 %	0.30	14min	59.4 %	0.32	2min:04s
uBERT_205	64.7 %	0.35	62.6 %	0.31	15min	62.0 %	0.35	2min:09s
uBERT_300	64.7 %	0.35	63.0 %	0.31	17min	63.0 %	0.36	2min:23s
uBERT_408	64.0 %	0.33	63.2 %	0.32	19min	64.3 %	0.38	2min:42s
uBERT_510	64.6 %	0.35	63.0 %	0.31	21min	64.2 %	0.38	3min:08s

⁵MCC ranges from -1 to 1, where 1 indicates perfect prediction, 0 indicates no better than random chance, and -1 indicates total disagreement between prediction and observation.

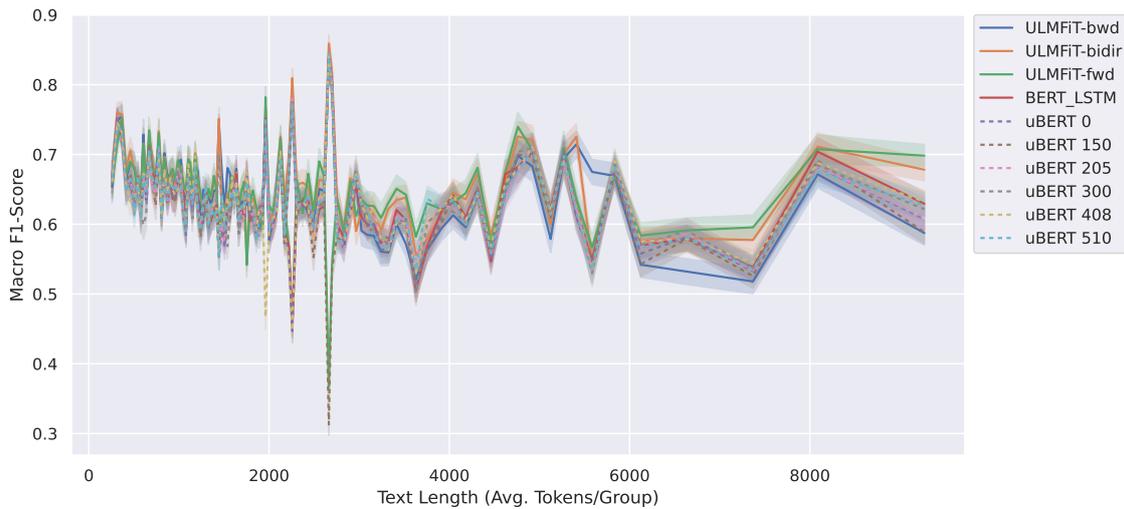


Figure 3: Macro-F1 score x Avg. Tokens/Group across different groups of same size ranked by the length. The error bars represent 95% confidence intervals obtained with bootstrap resampling.

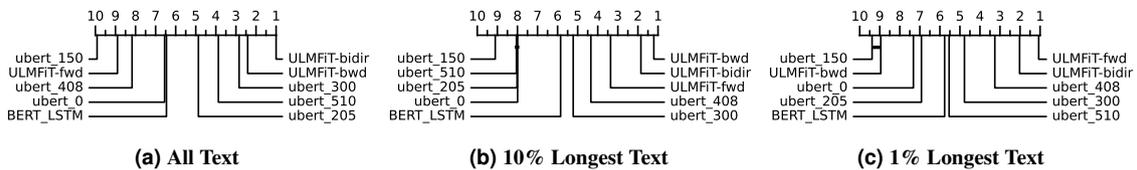


Figure 4: Critical difference diagram showing pairwise statistical comparison between baselines and varying overlap sizes for uBERT using the MCC. Connecting bars represent no statistical difference between methods.

Processing the Full Text Requires Overlap Comparing the BERT+LSTM baseline, which middle-truncates text when it exceeds input size, with our uBERT without overlap (uBERT₀), which uses the full text, we found that uBERT either underperformed or matched the baseline across all metrics. Notably, it performed worse on the 1% longest texts, where middle-truncation by BERT+LSTM occurs. **This suggests that merely processing the entire text is insufficient for longer inputs.** We hypothesize that non-overlapping chunks introduce noise due to abrupt segmentation, which degrades performance. Our results support this, showing that introducing overlap in uBERT configurations improves both Macro-F1 and MCC scores. The following uBERT configurations outperformed BERT+LSTM with statistical significance: uBERT₃₀₀, uBERT₅₁₀ and uBERT₂₀₅ on full test set; uBERT₄₀₈ and uBERT₃₀₀ on 10% longest; and uBERT₄₀₈, uBERT₃₀₀ and uBERT₅₁₀ on 1% longest. Thus, **incorporating overlap is crucial for maintaining semantic consistency and improving performance on longer texts.**

uBERT with Overlap is still Significantly Faster than ULMFiT As expected, introducing overlap in the uBERT architecture increased computational time overhead. However, across the full dataset and the 10% longest texts, uBERT configurations delivered better results than the BERT+LSTM baseline with comparable inference times. Notably, uBERT₄₀₈ achieved a 4x faster inference than ULMFiT on the 10% longest texts. For the 1% longest texts, the increased length required two passes of uBERT₄₀₈⁶, resulting

⁶With zero overlap, uBERT can process a maximum of 7,650 tokens in a single encoder pass. This limit arises because uBERT handles up to 15 chunks of 510 tokens each (excluding special tokens [CLS] and [SEP]). Therefore, documents longer than 7,650 tokens require at least two encoder passes.

in 1.8x slower inference compared to BERT+LSTM, which needed to middle-truncate in all cases. Despite this, `uBERT_408` slightly outperformed BERT+LSTM, narrowing the performance gap with ULMFiT while maintaining a faster inference, highlighting the efficiency and effectiveness of our approach. In summary, in all subsets, **uBERT configurations reduced the BERT+LSTM gap being significantly faster than ULMFiT.**

ULMFiT Outperforms uBERT on Longer Texts As shown in Figure 3, model differences become more clear with increasing text length. Big Bird consistently underperforms on longer texts, which is why it was excluded from the comparison charts. **While some uBERT configurations outperform BERT+LSTM on longer texts, F1 scores in both models degrade compared to full test dataset performance.** In contrast, ULMFiT models improve on longer texts compared to the full dataset. This suggests that **our architecture mitigates the degradation for longer text that is inherent to the BERT+LSTM approach, but still falls short of ULMFiT models, that handle better longer text but at a cost of 4x slower inference time.**

6. Conclusion and Future Work

Our experiments demonstrate that the `uBERT` model improves the handling of legal texts compared to baseline encoder-based models, particularly on longer texts, due to its capability to process entire documents using overlapping chunks. Despite the increased computational overhead, `uBERT` remains faster than ULMFiT. `uBERT` slightly outperforms BERT+LSTM, but still falls short of ULMFiT. Thus, further refinement is needed to fully match ULMFiT’s performance. Notably, even ULMFiT, the top-performing model in our experiments, achieves relatively low Macro-F1 scores, suggesting that processing the full text alone is insufficient for high performance on this task. In this direction, future research should expand the evaluation methodology by analyzing correctly and incorrectly classified cases across all tested models to assess whether specific characteristics of judicial decisions make them more prone to misclassification by certain models. Such an analysis, however, requires a multidisciplinary approach, including expert input from highly skilled legal practitioners.

Future research should also explore different chunking strategies to enhance text processing. Comparing syntactic chunking, which is based on grammatical structure, with semantic chunking, which is based on content meaning, could provide valuable insights. As this study focuses on a Portuguese-language dataset, evaluating these chunking approaches across datasets in multiple languages would help determine if optimal chunking strategies vary with language, contributing to more robust long-text segmentation and model performance across diverse linguistic contexts.

Acknowledgements

This work was supported by *CAPES* (Finance Code 001), *CNPQ* (grant 312360/23-1), *Programa Unificado de Bolsas de Estudo para Apoio à Formação de Estudantes (PUB-USP)*, *USP-IBM-FAPESP Center for Artificial Intelligence* (FAPESP grant 2019/07665-4), and Secretaria da Fazenda do Estado do Rio Grande do Sul (SEFAZ-RS), Brazil.

References

- Beltagy, I., Peters, M. E., and Cohan, A. (2020). Longformer: The long-document transformer.
- CNJ (2024). Conselho nacional de justiça - cnj. Accessed: 2024-08-05.
- Cui, J., Shen, X., Nie, F., Wang, Z., Wang, J., and Chen, Y. (2022). A survey on legal judgment prediction: Datasets, metrics, models and challenges. *arXiv preprint arXiv:2204.04859v1*.
- Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7:1–30.
- Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805v2*.
- Ferraz, T. P., Alcoforado, A., Bustos, E., Oliveira, A. S., Gerber, R., Müller, N., d’Almeida, A. C., Veloso, B. M., and Costa, A. H. R. (2021). Debacer: a method for slicing moderated debates. In *Anais do XVIII Encontro Nacional de Inteligência Artificial e Computacional*, pages 667–678. Sociedade Brasileira de Computação-SBC.
- Guu, K., Lee, K., Tung, Z., Pasupat, P., and Chang, M.-W. (2020). Realm: Retrieval-augmented language model pre-training.
- Hasan, M. (2022). Transformers in natural language processing.
- Hoang, T. D., Bui, C. M., and Bui, N. (2023). Viettel-AI at SemEval-2023 task 6: Legal document understanding with longformer for court judgment prediction with explanation. In Ojha, A. K., Doğruöz, A. S., Da San Martino, G., Tayyar Madabushi, H., Kumar, R., and Sartori, E., editors, *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 862–868, Toronto, Canada. Association for Computational Linguistics.
- Howard, J. and Ruder, S. (2018). Fine-tuned language models for text classification. *CoRR*, abs/1801.06146.
- Katz, D., Hartung, D., Gerlach, L., Jana, A., and Bommarito, M. (2023). Natural language processing in the legal domain. *arXiv preprint arXiv:2302.12039v1*.
- Li, Q., Peng, H., Li, J., Xia, C., Yang, R., Sun, L., Yu, P. S., and He, L. (2022). A survey on text classification: From traditional to deep learning. *ACM Trans. Intell. Syst. Technol.*, 13(2).
- Mahari, R., Stambach, D., Ash, E., and Pentland, A. (2023). The law and NLP: Bridging disciplinary disconnects. In Bouamor, H., Pino, J., and Bali, K., editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3445–3454, Singapore. Association for Computational Linguistics.
- Malik, V., Sanjay, R., Nigam, S. K., Ghosh, K., Guha, S. K., Bhattacharya, A., and Modi, A. (2021). ILDC for CJPE: Indian legal documents corpus for court judgment prediction and explanation. In Zong, C., Xia, F., Li, W., and Navigli, R., editors, *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)*, pages 4046–4062, Online. Association for Computational Linguistics.

- Medvedeva, M. and McBride, P. (2023). Legal judgment prediction: If you are going to do it, do it right. In Preotiuc-Pietro, D., Goanta, C., Chalkidis, I., Barrett, L., Spanakis, G., and Aletras, N., editors, *Proceedings of the Natural Legal Language Processing Workshop 2023*, pages 73–84, Singapore. Association for Computational Linguistics.
- Menezes-Neto, E. J. d. and Clementino, M. B. M. (2022). Using deep learning to predict outcomes of legal appeals better than human experts: A study with data from brazilian federal courts. *PLOS ONE*, 17(7):1–20.
- Pappagari, R., Żelasko, P., Villalba, J., Carmiel, Y., and Dehak, N. (2019). Hierarchical transformers for long document classification.
- Tenney, I., Das, D., and Pavlick, E. (2019). Bert rediscovers the classical nlp pipeline.
- Wan, L., Seddon, M., Papageorgiou, G., and Bernardoni, M. (2019). Long-length legal document classification. *arXiv preprint arXiv:1912.06905v1*.
- Zaheer, M., Guruganesh, G., Dubey, A., Ainslie, J., Alberti, C., Ontañón, S., Pham, P., Ravula, A., Wang, Q., Yang, L., and Ahmed, A. (2020). Big bird: Transformers for longer sequences. *CoRR*, abs/2007.14062.
- Zhang, L., Wang, W., Yu, K., huang, J., Lyu, Q., Xue, H., and Hetang, C. (2023). Sliding-bert: Striding towards conversational machine comprehension in long context. *Adv. Artif. Intell. Mach. Learn.*, 3:1325–1339.
- Zhu, H., Mak, D., Gioannini, J., and Xia, F. (2020). NLPStatTest: A toolkit for comparing NLP system performance. In Wong, D. and Kiela, D., editors, *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 40–46, Suzhou, China. Association for Computational Linguistics.

Syntactic parsing: where are we going?

Lucelene Lopes¹, Thiago Alexandre Salgueiro Pardo¹, Magali S. Duran¹

¹Núcleo Interinstitucional de Linguística Computacional (NILC)
Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo
São Carlos-SP, Brazil

***Abstract.** In this review & opinion paper, we discuss the options and challenges for syntactic parsing. Despite significant advances in recent years, driven primarily by neural network architectures, parsing accuracy appears to be approaching a plateau. This paper proposes a reflection on the factors that may possibly be influencing such results and suggests some future paths.*

Motivation

The importance of good part of speech tagging and parsing annotation tools for downstream Natural Language Processing (NLP) tasks is acknowledged by several publications in the history of the area, including both more classic (symbolic and statistic) approaches and new (usually neural-based) ones. In particular, the rise of “Universal Dependencies” (UD) framework¹ [Nivre et al. 2016, de Marneffe et al. 2021] has sparked renewed interest in dependency parsing, driving new efforts in syntax studies and parsing in NLP.

This review & opinion paper attempts to draw a landscape of more recent parsing efforts that align to UD standards, trying to figure out the potential limits of the task with current methods and what other strategies might be adopted for keeping improving the achieved results in the area. Such initiative is bold and naturally subject to failure, as natural languages have diverse characteristics and there are always new NLP methods emerging. Knowing this, this article makes a selection of works from the literature, choosing relatively recent and widely cited approaches in the area in order to draw some tentative (and certainly temporally anchored) conclusions.

Besides the possibly interesting work selection and overview that supported this paper, our contribution includes an exercise of “keeping the head above water”, showing how far we have come and the imperfections of the landscape.

On current parsing techniques

The use of neural networks for detection of patterns, and consequently, the prediction of part of speech tags and dependency relations became the preferred method in the area [Goldberg 2016]. Within neural networks, several techniques as Long Short-Term Memory (LSTM) in its various versions [Van Houdt et al. 2020], together with other deep learning techniques [Dozat and Manning 2016], have been employed in the last decade with consistent advances for well resourced languages. The latest evolution brought by the self-attention methods [Vaswani et al. 2017], based on the famous Transformers, goes back a few years now, but it is still one of the main reasons for recent improvements.

¹<https://universaldependencies.org/>

Overall, although different criteria could be used, in this paper we distinguish the parsing efforts according to the generic parsing tools or specific language parsing initiatives; and basic technology employed (e.g., BiLSTM, Deep Biaffine, and Self-Attention).

The more popular parsing tools, within UD standard, are the UDPipe in its versions 1.3 and 2.0 [Straka et al. 2016, Straka 2018], Stanza pipeline [Qi et al. 2020], UDify [Kondratyuk and Straka 2019], and AllenNLP pipeline [Dozat and Manning 2016]. Other less popular tools were developed, but apparently had fewer number of users, as the Diaparser [Attardi et al. 2021], UDapter [Üstün et al. 2020], UUParser [de Lhoneux et al. 2017], LAL-Parser [Mrini et al. 2019], and Hierarchical Pointer Network algorithm [Fernández-González and Gómez-Rodríguez 2023].

These parsers usually focus their efforts to cover several languages, being clearly multilingual. Some of these tools were specifically designed to cover the large set of languages available at the UD repository (which currently includes over 150 languages). However, from a technological point of view, the tools have considerable differences, although all of them make use of neural network models.

The technology of Bidirectional LSTM (BiLSTM) [Van Houdt et al. 2020] is frequently employed by many systems, including UDPipe 2.0, Stanza, and Hierarchical Pointer Networks algorithm. The Deep Biaffine technology [Dozat and Manning 2016] is found in AllenNLP pipeline, but also in tools as Diaparser and UDapter. Self-attention [Vaswani et al. 2017] is found in LAL-Parser and UDify tools. Additionally, the mentioned tools show differences on offering a static model or the possibility to perform model construction through a training set and/or to adopt pre-trained word embeddings.

Parsing results

The best values reported for each of the previously cited parsing methods are shown in Table 1. We chose to report only the Label Attachment Score (LAS), as this is usually the most adopted evaluation metric and also one of the most punitive metrics, as it measures the accuracy of the dependency relation identification and the tokens related as head and dependent. The table also indicates the language for which the highest LAS was reported.

Table 1. Highest LAS reported for the generic parsing tools.

parsing system	highest LAS	language	cited technology	publication
UUParser	87.34%	Portuguese	BiLSTM	2017
Stanza	90.01%	Spanish	Deep Biaffine	2020
UDPipe 1.3	91.20%	Hindi	NN Classifier	2016
UDapter	92.60%	Italian	Deep Biaffine	2020
Diaparser	93.65%	Italian	Deep Biaffine	2021
UDify	93.70%	Russian	Self-Attention	2019
UDPipe 2.0	94.53%	Russian	BiLSTM	2018
AllenNLP pipeline	94.60%	English	Deep Biaffine	2016
Hier. Pointer Networks	96.15%	English	BiLSTM	2023
LAL-parser	96.29%	English	Self-Attention	2019

The performance of the parsing methods vary considerably according to the language to which they are applied, as the scientific literature has shown. For example,

for UDPipe 2, the reported LAS for Spanish and Italian can be as low as 80.68% and 77.34%, respectively. For AllenNLP pipeline, LAS for Chinese and Spanish was 85.38% and 91.65%, respectively. The values shown in the table may also reflect the number of tested languages. While UDify and UDPipe test over more than 70 languages, AllenNLP pipeline, UUParser, and LAL-parser test for only 6, 5, and 2 languages, respectively.

Focusing only on the highest LAS accuracy as presented in Table 1, it is noticeable that the majority of the highest scores are over 90% of accuracy. These numbers suggest that the State Of The Art (SOTA) for LAS is attainable despite of the technology employed, date of publication, and even specificity of each parsing development. Observing the three best reported results, we see different techniques and that English shows the best scores (probably because English is the best resourced language).

This fact suggests that, after the spread of neural network-based models, the quality of the training model plays a more important role than the specific technology employed. As such, the variations for different languages seem to reflect the quality of the training data for each language. For example, LAS for UDify for a low resourced language as Breton is as low as 40.19%, which is much lower than the 93.70% maximum attained for Russian.

Fortunately, the literature is abundant in terms of efforts for specific languages. These works usually are presented either with the construction of a specific corpus for the target language, or transferring learning from a better resourced language towards the low resourced one. Observing the works dedicated to specific languages, we found a reasonable number of publications, some of which are summarized in Table 2.

Table 2. Highest LAS reported by specific language efforts.

work	LAS	language	overall approach
[Dione 2021]	31.43%	Yoruba	Transfer learning
[Brigada Villa and Giarda 2023]	58.70%	Old English	Transfer learning
[Cassidy et al. 2022]	59.34%	Indonesian	Transfer learning
[Lusito and Maillard 2021]	60.74%	Ligurian	Corpus building
[Baig et al. 2021]	62.90%	Urdu	Corpus building
[Dione 2021]	67.83%	Wolof	Transfer learning
[Türk et al. 2022]	76.04%	Turkish	Corpus building
[Ghiffari et al. 2023]	79.22%	Irish	Corpus building
[Pedrazzini and Eckhoff 2021]	79.66%	Old Slavic	Transfer learning
[Sánchez-Rodríguez et al. 2024]	84.31%	Galician	Corpus building
[Alves et al. 2021]	89.09%	Croatian	Transfer learning
[Branco et al. 2022]	92.54%	Portuguese	Corpus building
[Kabiri et al. 2022]	92.68%	Persian	Corpus building
[Gamba and Zeman 2023]	94.61%	Latin	Corpus building
[Lopes and Pardo 2024]	94.70%	Portuguese	Corpus building

The examples summarized in Table 2 show efforts that can be grouped into attempts to serve very low resourced languages (as Old English, Old Slavic, Ligurian, Urdu, Bambara, Wolof, and Indonesian) and low resourced languages (as Turkish, Croatian, Galician, Irish, Persian, Latin, and Portuguese). While the very low resourced languages

attempts are mostly based on transfer learning, the languages better resourced mostly center the efforts in building better corpora to be used to train specific models.

The observation of LAS in Table 2 shows that the best reported results are also above the 90% score of the generic parsing methods (Table 1). Obviously, the hard cases, as Yoruba and Old English, show low accuracy despite the efforts, probably because they are low-resourced languages. However, it is noticeable the accuracy achieved by transfer learning for Old Slavic and Croatian, as well as the high values for Persian, Latin, and Portuguese with the production of high quality training corpora.

Where can we head to?

The advent of popular neural network methods in the last decade has brought impressive progress in several areas of NLP, bringing Artificial Intelligence to the center of topics in all areas of the human knowledge. For parsing tasks, specifically, using UD standards, we notice the increase of quality since 2016. However, improvements seem to reach a limit up to 96% accuracy, and it is noticeable that no specificity show a clear predominance.

It is also well known that languages with few resources may not be able to benefit from the advantages of SOTA methods. It would be better for these languages to invest in more classic methods or in the improvement of resources through corpora building including careful annotation. Specific techniques like data augmentation and joint task resolution may also be interesting ways (see, e.g., the work of [Yshaayahu Levi and Tsarfaty 2024] for Hebrew parsing). Such paths may also be relevant for languages already reaching accuracy around 95%, i.e., already delivering SOTA results.

Another relevant question is if the search for a better accuracy (over 96%) is a realistic goal. Should we make our peace with these missing 4% due to a natural inaccuracy of dependency annotation? Looking at the best method for a specific language (Portuguese), the authors [Lopes and Pardo 2024] [Duran et al. 2023a] [Duran et al. 2023b] discuss some reasons for the remaining errors that are also cited in the literature: under-represented phenomena in the training corpus (that might be solved by data augmentation and/or more corpus annotation) and difficult annotation issues (as to decide which is the head of a prepositional phrase) that sometimes may challenge even the humans. Personally, we believe that the above 99% accuracy already achieved for part of speech tagging may be achieved for parsing too. However, it may require to simplify some syntactic distinctions or to look for new approaches to the parsing problem.

The interested reader may find more information at the POeTiSA project web portal: <https://sites.google.com/icmc.usp.br/poetisa>

Acknowledgements

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. The project was also supported by 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.

References

- Alves, D., Bekavac, B., and Tadić, M. (2021). Typological approach to improve dependency parsing for Croatian language. In Dakota, D., Evang, K., and Kübler, S., editors, *Proceedings of the 20th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest)*, pages 1–11, Sofia, Bulgaria. Association for Computational Linguistics.
- Attardi, G., Sartiano, D., and Simi, M. (2021). Biaffine dependency and semantic graph parsing for Enhanced Universal dependencies. In Oepen, S., Sagae, K., Tsarfaty, R., Bouma, G., Seddah, D., and Zeman, D., editors, *Proceedings of the 17th International Conference on Parsing Technologies and the IWPT 2021 Shared Task on Parsing into Enhanced Universal Dependencies*, pages 184–188, Online. Association for Computational Linguistics.
- Baig, A., Rahman, M. U., Shah, A. S., and Abbasi, S. (2021). Universal dependencies for urdu noisy text. *International Journal of Advanced Trends in Computer Science and Engineering*.
- Branco, A., Silva, J. R., Gomes, L., and António Rodrigues, J. (2022). Universal grammatical dependencies for Portuguese with CINTIL data, LX processing and CLARIN support. In Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Odijk, J., and Piperidis, S., editors, *Proceedings of the Thirteenth Language Resources and Evaluation Conference (LREC)*, pages 5617–5626, Marseille, France. European Language Resources Association.
- Brigada Villa, L. and Giarda, M. (2023). Using modern languages to parse ancient ones: a test on Old English. In Beinborn, L., Goswami, K., Muradoğlu, S., Sorokin, A., Kumar, R., Shcherbakov, A., Ponti, E. M., Cotterell, R., and Vylomova, E., editors, *Proceedings of the 5th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*, pages 30–41, Dubrovnik, Croatia. Association for Computational Linguistics.
- Cassidy, L., Lynn, T., Barry, J., and Foster, J. (2022). TwittIrish: A Universal Dependencies treebank of tweets in Modern Irish. In Muresan, S., Nakov, P., and Villavicencio, A., editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6869–6884, Dublin, Ireland. Association for Computational Linguistics.
- de Lhoneux, M., Stymne, S., and Nivre, J. (2017). Arc-hybrid non-projective dependency parsing with a static-dynamic oracle. In Miyao, Y. and Sagae, K., editors, *Proceedings of the 15th International Conference on Parsing Technologies*, pages 99–104, Pisa, Italy. Association for Computational Linguistics.
- de Marneffe, M.-C., Manning, C. D., Nivre, J., and Zeman, D. (2021). Universal Dependencies. *Computational Linguistics*, 47(2):255–308.
- Dione, C. M. B. (2021). Multilingual dependency parsing for low-resource African languages: Case studies on Bambara, Wolof, and Yoruba. In Oepen, S., Sagae, K., Tsarfaty, R., Bouma, G., Seddah, D., and Zeman, D., editors, *Proceedings of the 17th International Conference on Parsing Technologies and the IWPT 2021 Shared Task on*

- Parsing into Enhanced Universal Dependencies*, pages 84–92, Online. Association for Computational Linguistics.
- Dozat, T. and Manning, C. D. (2016). Deep biaffine attention for neural dependency parsing. *CoRR*, abs/1611.01734.
- Duran, M., das Graças Nunes, M., and Pardo, T. A. (2023a). Construções sintáticas do português que desafiam a tarefa de parsing: uma análise qualitativa. In *Anais do XIV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 424–433, Porto Alegre, RS, Brasil. SBC.
- Duran, M. S., Nunes, M. d. G. V., and Pardo, T. A. S. (2023b). Avaliação qualitativa do analisador sintático udpipe 2 treinado sobre o corpus jornalístico porttinari-base. Technical report, Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo.
- Fernández-González, D. and Gómez-Rodríguez, C. (2023). Dependency parsing with bottom-up hierarchical pointer networks. *Information Fusion*, 91:494–503.
- Gamba, F. and Zeman, D. (2023). Universalising Latin Universal Dependencies: a harmonisation of Latin treebanks in UD. In Grobol, L. and Tyers, F., editors, *Proceedings of the Sixth Workshop on Universal Dependencies (UDW, GURT/SyntaxFest)*, pages 7–16, Washington, D.C. Association for Computational Linguistics.
- Ghiffari, F. A. A., Alfina, I., and Azizah, K. (2023). Cross-lingual transfer learning for Javanese dependency parsing. In Li, D., Mahendra, R., Tang, Z. P., Jang, H., Murawaki, Y., and Wong, D. F., editors, *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 1–9, Nusa Dua, Bali. Association for Computational Linguistics.
- Goldberg, Y. (2016). A primer on neural network models for natural language processing. *J. Artif. Int. Res.*, 57(1):345–420.
- Kabiri, R., Karimi, S., and Surdeanu, M. (2022). Informal Persian Universal Dependency treebank. In Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Odijk, J., and Piperidis, S., editors, *Proceedings of the Thirteenth Language Resources and Evaluation Conference (LREC)*, pages 7096–7105, Marseille, France. European Language Resources Association.
- Kondratyuk, D. and Straka, M. (2019). 75 languages, 1 model: Parsing Universal Dependencies universally. In Inui, K., Jiang, J., Ng, V., and Wan, X., editors, *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 2779–2795, Hong Kong, China. Association for Computational Linguistics.
- Lopes, L. and Pardo, T. (2024). Towards portparser - a highly accurate parsing system for Brazilian Portuguese following the Universal Dependencies framework. In Gamallo, P., Claro, D., Teixeira, A., Real, L., Garcia, M., Oliveira, H. G., and Amaro, R., editors, *Proceedings of the 16th International Conference on Computational Processing*

- of Portuguese - Vol. 1*, pages 401–410, Santiago de Compostela, Galicia/Spain. Association for Computational Linguistics.
- Lusito, S. and Maillard, J. (2021). A Universal Dependencies corpus for Ligurian. In de Lhoneux, M. and Tsarfaty, R., editors, *Proceedings of the Fifth Workshop on Universal Dependencies (UDW, SyntaxFest)*, pages 121–128, Sofia, Bulgaria. Association for Computational Linguistics.
- Mrini, K., Deroncourt, F., Bui, T., Chang, W., and Nakashole, N. (2019). Rethinking self-attention: An interpretable self-attentive encoder-decoder parser. *CoRR*, abs/1911.03875.
- Nivre, J., de Marneffe, M.-C., Ginter, F., Goldberg, Y., Hajič, J., Manning, C. D., McDonald, R., Petrov, S., Pyysalo, S., Silveira, N., Tsarfaty, R., and Zeman, D. (2016). Universal Dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC)*, pages 1659–1666, Portorož, Slovenia. ELRA.
- Pedrazzini, N. and Eckhoff, H. M. (2021). Oldslavnet: A scalable early slavonic dependency parser trained on modern language data. *Software Impacts*, 8:100063.
- Qi, P., Zhang, Y., Zhang, Y., Bolton, J., and Manning, C. D. (2020). Stanza: A python natural language processing toolkit for many human languages. In Celikyilmaz, A. and Wen, T.-H., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.
- Sánchez-Rodríguez, X., Sarymsakova, A., Castro, L., and Garcia, M. (2024). Increasing manually annotated resources for Galician: the parallel Universal Dependencies treebank. In Gamallo, P., Claro, D., Teixeira, A., Real, L., Garcia, M., Oliveira, H. G., and Amaro, R., editors, *Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 1*, pages 587–592, Santiago de Compostela, Galicia/Spain. Association for Computational Linguistics.
- Straka, M. (2018). UDPipe 2.0 prototype at CoNLL 2018 UD shared task. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 197–207.
- Straka, M., Hajič, J., and Straková, J. (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)*, pages 4290–4297, Portorož, Slovenia. European Language Resources Association (ELRA).
- Türk, U., Atmaca, F., Özateş, c. B., Berk, G., Bedir, S. T., Köksal, A., Başaran, B. O., Güngör, T., and Özgür, A. (2022). Resources for turkish dependency parsing: introducing the boun treebank and the boat annotation tool. *Lang. Resour. Eval.*, 56(1):259–307.
- Üstün, A., Bisazza, A., Bouma, G., and van Noord, G. (2020). UDapter: Language adaptation for truly Universal Dependency parsing. In Webber, B., Cohn, T., He, Y., and Liu, Y., editors, *Proceedings of the 2020 Conference on Empirical Methods in*

- Natural Language Processing (EMNLP)*, pages 2302–2315, Online. Association for Computational Linguistics.
- Van Houdt, G., Mosquera, C., and Nápoles, G. (2020). A review on the long short-term memory model. *Artificial Intelligence Review*, 53.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. (2017). Attention is all you need. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.
- Yshaayahu Levi, D. and Tsarfaty, R. (2024). A truly joint neural architecture for segmentation and parsing. In Graham, Y. and Purver, M., editors, *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1408–1420, St. Julian’s, Malta. Association for Computational Linguistics.

Segmentação Textual Baseada em Tópicos em Português Utilizando BERTimbau

Luciano A. C. da Silva¹, Maiara S. F. Rodrigues¹, Adriana P. Archanjo¹, Luis Pessoa¹, Miguel L. Silva¹, Thiago F. de Almeida¹, Leonardo Silveira²,

¹CPQD - Centro de Pesquisa e Desenvolvimento, Campinas, SP, Brasil

²Pontifícia Universidade Católica de Campinas, SP, Brasil

luciano.augusto.silva@usp.br, maiara.frodrigues2000@gmail.com, prestoarch@hotmail.com, {luisp, mfilho, tfelipea}@cpqd.com.br, leonardo.silveira@ga.ita.br

Abstract. *In this work, we explore text segmentation for Portuguese using the BERTimbau model, with datasets derived from machine translation and online news sources. We obtained $P_k = 6.89$ for an in-domain evaluation, but worse results in out-of-domain evaluations, highlighting the importance of a diverse training set to improve generalization across multiple domains.*

Resumo. *Neste trabalho, exploramos a segmentação textual para o português utilizando o modelo BERTimbau, com bases de dados construídas usando tradução automática e a partir de notícias online. Obtivemos $P_k = 6,89$ para uma avaliação dentro do domínio, mas resultados piores em avaliações fora do domínio, destacando a importância de uma base de treinamento diversificada para melhorar a generalização em múltiplos domínios.*

1. Introdução

Com o aumento na geração de conteúdo textual não estruturado, como transcrições automáticas de notícias, aulas e reuniões, há também um crescente interesse em extrair de forma eficiente informações relevantes desse material [Retkowski and Waibel 2024, Gklezakos et al. 2024]. Por exemplo, pode ser desafiador encontrar o início de um determinado tópico discutido na transcrição de uma longa reunião, a menos que essa transcrição esteja devidamente estruturada. A segmentação textual baseada em tópicos é uma tarefa de Processamento de Linguagem Natural (PLN) que divide um texto longo em segmentos não sobrepostos, de acordo com as mudanças de tópico [Hearst 1997]. Essa ferramenta permite estruturar e compreender melhor grandes volumes de dados, facilitando a busca e a extração de informações.

Há poucos trabalhos recentes sobre segmentação textual em português [Cardoso et al. 2017, Francisco 2018]. Neste artigo, exploramos a segmentação textual baseada em tópicos para o português, aplicando a abordagem proposta em [Yu et al. 2023], utilizando o modelo BERTimbau [Souza et al. 2023]. Construímos os conjuntos de dados de treinamento e teste por meio de tradução automática para o português, e utilizando notícias extraídas da internet.

2. Metodologia

Neste trabalho, utilizamos a abordagem proposta por [Yu et al. 2023] que trata a segmentação textual como um problema de classificação de uma sequência de sentenças,

em que se deseja identificar a última sentença de cada tópico, ou seja, identificar as fronteiras dos segmentos. O componente principal é um modelo de linguagem pré-treinado do tipo *Transformer encoder* [Vaswani et al. 2023], que produz a representação contextual das sentenças do texto de entrada. Cada representação de sentença é usada na classificação de fronteira do segmento, conforme mostrado na Figura 1.

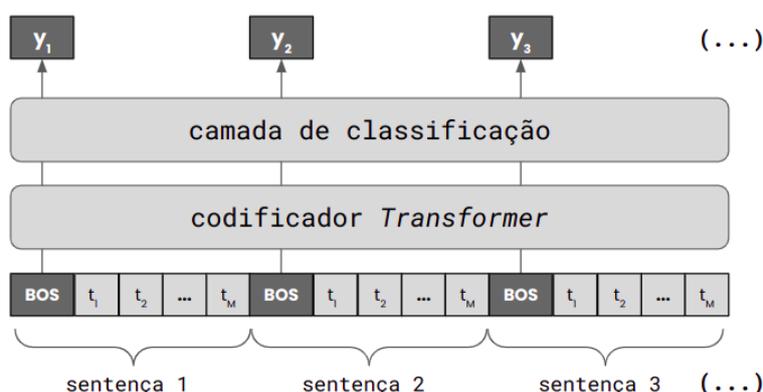


Figura 1. Estrutura do modelo de segmentação proposto por [Yu et al. 2023]

Em [Yu et al. 2023], além da tarefa principal de segmentação baseada em tópicos, são definidas duas tarefas auxiliares adicionais, *Topic-aware Sentence Structure Prediction* (TSSP) e *Contrastive Semantic Similarity Learning* (CSSL), com o objetivo de modelar a coerência textual e obter melhores resultados na segmentação. O modelo é treinado de forma supervisionada, otimizando a soma das perdas das três tarefas definidas, sobre um conjunto de treinamento devidamente anotado.

Neste trabalho, utilizamos *datasets* para o treinamento e a avaliação obtidos por meio de tradução automática para português ou construídos a partir de notícias em português extraídas da internet. Os *datasets* WikiSection e WIKI-50 foram usados por [Yu et al. 2023] e passaram pelo processo de tradução automática usando a *API* de tradução da Google. O *dataset* WikiSection [Arnold et al. 2019] foi usado para treinamento e avaliação, e consiste num conjunto de 38K artigos em inglês e alemão, nos domínios de doenças e cidades. Após a tradução, restaram 3.590 documentos no domínio de doenças e 19.539 documentos no domínio de cidades. O *dataset* WIKI-50 [Koshorek et al. 2018] foi usado apenas para avaliação, e consiste originalmente em um conjunto de 50 amostras em inglês, provenientes da Wikipedia.

Para a avaliação dos modelos, utilizamos também *datasets* em português construídos a partir de notícias extraídas com *webscraping* do portal G1¹ (portal de notícias do Grupo Globo de Comunicação), e do canal de notícias do IBGE² (Instituto Brasileiro de Geografia e Estatística). Os documentos de texto foram formados pela concatenação aleatória de notícias, sendo cada notícia considerada um segmento de tópico diferente. No caso do *dataset* G1, foram gerados 454 documentos a partir de 1.300 notícias. Para o *dataset* IBGE, foram gerados 1.517 documentos a partir de 3.376 notícias.

¹<https://g1.globo.com/tecnologia/noticia/2012/11/siga-o-g1-por-rss.html>

²<https://servicodados.ibge.gov.br/api/docs/noticias?versao=3>

Como o nosso objetivo é aplicar a segmentação para o português, substituímos o modelo usado em [Yu et al. 2023] pelo modelo BERTimbau [Souza et al. 2023], pré-treinado para o português do Brasil. Utilizamos as versões BERTimbau Base (110M de parâmetros) e BERTimbau Large (335M de parâmetros)³.

O treinamento foi realizado em uma GPU NVIDIA T4, usando BERTimbau Base e Large, com 70% do *dataset* WikiSection em português, por 5 épocas, com *learning rate* de 5×10^{-5} , *batch size* de 2 e gradiente acumulado de 2. Criamos sempre um modelo treinado com WikiSection/cidades e o outro modelo treinado com WikiSection/doenças. No caso do BERTimbau Large, o treinamento durou aproximadamente 2 dias e 5 horas para o conjunto de cidades e pouco mais de 11 horas para o conjunto de doenças.

A avaliação dos modelos seguiu a mesma linha de [Yu et al. 2023]. Usamos três métricas usuais para avaliação de segmentação textual: F_1 , P_k [Beeferman et al. 1999], e *WindowDiff* [Pevzner and Hearst 2002]. No caso das métricas P_k e *WindowDiff*, quanto menor o valor, melhor o desempenho. No caso da métrica F_1 , quanto maior o valor, melhor o desempenho. A avaliação dentro do domínio de treinamento foi realizada com 20% do *dataset* WikiSection em português. Os *datasets* WIKI-50, G1 e IBGE são usados apenas para avaliação fora do domínio de treinamento.

3. Resultados

As Tabelas 1 e 2 apresentam os resultados de avaliação dos modelos usando BERTimbau, criados e avaliados para o português, dentro do mesmo domínio, com os *datasets* WikiSection/cidades e WikiSection/doenças. Também são apresentados os resultados para o inglês correspondentes ao modelo BERT Base [Devlin et al. 2018], obtidos por [Yu et al. 2023].

Modelo	F_1	P_k	WD
(en) BERT Base [Yu et al. 2023]	80,16	8,22	10,19
(pt) BERTimbau Base	87,41	7,07	8,55
(pt) BERTimbau Large	87,59	6,89	8,37

Tabela 1. Resultados dos modelos criados e avaliados com o *dataset* WikiSection / cidades. BERT Base avaliado em inglês, BERTimbau em português.

Modelo	F_1	P_k	WD
(en) BERT Base [Yu et al. 2023]	68,26	18,29	22,06
(pt) BERTimbau Base	76,91	17,16	19,45
(pt) BERTimbau Large	77,77	16,55	18,76

Tabela 2. Resultados dos modelos criados e avaliados com o *dataset* WikiSection / doenças. BERT Base avaliado em inglês, BERTimbau em português.

As métricas de avaliação obtidas com os modelos BERTimbau para o português são melhores e próximas àquelas apresentadas por [Yu et al. 2023] em inglês. Neste caso, devemos considerar também que o modelo criado para o português usando BERTimbau Large é maior que o modelo usado em [Yu et al. 2023].

³<https://huggingface.co/neuralmind/bert-base-portuguese-cased>

A Tabela 3 apresenta os resultados da avaliação de dois modelos criados para o português nos domínios de cidades e doenças, usando o BERTimbau Large, e avaliados fora do domínio de treinamento, nos *datasets* WIKI-50, G1 e IBGE.

Dataset	Modelo / cidades			Modelo / doenças		
	F_1	P_k	WD	F_1	P_k	WD
Wiki50	15,43	35,01	35,36	12,97	35,98	36,02
G1	64,66	13,62	17,28	54,81	25,61	32,42
IBGE	43,12	20,12	21,06	43,36	23,40	26,55

Tabela 3. Avaliação fora do domínio de treinamento. Modelos com o BERTimbau Large criados com o *dataset* WikiSection/cidades e WikiSection/doenças.

O desempenho do modelo fora do domínio de treinamento foi inferior ao desempenho dentro do domínio. Os resultados foram melhores para o modelo treinado com o *dataset* WikiSection/cidades. De fato, segundo [Arnold et al. 2019], o conteúdo do *dataset* WikiSection apresenta características distintas para cada domínio: WikiSection/doenças é de domínio científico restrito com linguagem específica, enquanto WikiSection/cidades é de domínio geral mais diverso, mais próximo de um conteúdo de notícias. Isso sugere que a composição de dados de treinamento pode ajudar a obter um modelo para segmentação textual que generalize melhor para múltiplos domínios.

4. Conclusão

Neste trabalho, exploramos a segmentação textual para o português, seguindo a abordagem de [Yu et al. 2023], mas utilizando o modelo pré-treinado para o português BERTimbau [Souza et al. 2023]. Empregamos bases de treinamento e teste construídas usando a tradução automática de bases existentes, além de bases de teste construídas a partir de notícias em português recuperadas da internet. Obtivemos ótimos resultados na segmentação de texto dentro do mesmo domínio para o português, semelhante ao que foi obtido por [Yu et al. 2023] para o inglês. Nossos resultados sugerem a eficácia do método empregado para a criação do modelo em português e a importância de usar uma base de treinamento de domínio diversificado para obter um modelo que generalize melhor para múltiplos domínios.

Para trabalhos futuros, pretendemos explorar modelos diferentes e buscar uma composição mais variada de dados de treinamento para obter um modelo que generalize melhor para vários domínios. Além disso, desejamos estudar a segmentação textual de transcrições automáticas obtidas com reconhecimento de fala, e explorar a segmentação de textos muito longos, considerando a típica limitação do contexto de entrada de modelos baseados em *Transformer* [Vaswani et al. 2023].

Agradecimentos

Este projeto foi apoiado pelo Ministério da Ciência, Tecnologia e Inovações, com recursos da Lei no 8.248, de 23 de outubro de 1991, no âmbito do PPI-SOFTEX, coordenado pela Softex e publicado PDI 03, DOU 01245.023862/2022-14.

Referências

- Arnold, S., Schneider, R., Cudré-Mauroux, P., Gers, F. A., and Löser, A. (2019). Sector: A neural model for coherent topic segmentation and classification. *Transactions of the Association for Computational Linguistics*, 7:169–184.
- Beeferman, D., Berger, A. L., and Lafferty, J. D. (1999). Statistical models for text segmentation. *Machine Learning*, 34:177–210.
- Cardoso, P. C., Pardo, T. A., and Taboada, M. (2017). Subtopic annotation and automatic segmentation for news texts in brazilian portuguese. *Corpora*, 12(1):23–54.
- Devlin, J., Chang, M., Lee, K., and Toutanova, K. (2018). BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.
- Francisco, O. J. (2018). Recuperação de informação em atas de reunião utilizando segmentação textual e extração de tópicos. Dissertação de mestrado, Universidade Federal de São Carlos, Sorocaba.
- Gklezakos, D. C., Misiak, T., and Bishop, D. (2024). Treeseq: Hierarchical topic segmentation of large transcripts. *arXiv preprint arXiv:2407.12028*.
- Hearst, M. A. (1997). Text tiling: Segmenting text into multi-paragraph subtopic passages. *Computational linguistics*, 23(1):33–64.
- Koshorek, O., Cohen, A., Mor, N., Rotman, M., and Berant, J. (2018). Text segmentation as a supervised learning task. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 469–473.
- Pevzner, L. and Hearst, M. A. (2002). A critique and improvement of an evaluation metric for text segmentation. *Computational Linguistics*, 28(1):19–36.
- Retkowski, F. and Waibel, A. (2024). From text segmentation to smart chaptering: A novel benchmark for structuring video transcriptions. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 406–419.
- Souza, F., Nogueira, R., and Lotufo, R. (2023). Bert models for brazilian portuguese: Pre-training, evaluation and tokenization analysis. *Applied Soft Computing*, 149:110901.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2023). Attention is all you need.
- Yu, H., Deng, C., Zhang, Q., Liu, J., Chen, Q., and Wang, W. (2023). Improving long document topic segmentation models with enhanced coherence modeling. In Bouamor, H., Pino, J., and Bali, K., editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5592–5605, Singapore. Association for Computational Linguistics.

Avaliação de modelos para detecção de ataques de *replay* usando diferentes bases de dados

Giovana Y. Nakashima¹, Higor D. C. Santos¹, Jone W. M. Soares¹,
Mário Uliani Neto¹, Fernando O. Runstein¹, Ricardo P. V. Violato¹, Marcus Lima²

¹CPQD - Centro de Pesquisa e Desenvolvimento, Campinas, SP, Brasil

²Pontifícia Universidade Católica de Campinas, SP, Brasil

{giovana.nakashima, higorcea, jonewisney, marcuslima3}@gmail.com
{uliani, runstein, rviolato}@cpqd.com.br

Abstract. *A replay attack is a speech forgery used in an attempt to authenticate a speaker. Deep neural networks have been proposed as methods for detecting fraudulent audio. In view of the use of these models in real applications, in addition to good learning performance it is expected that the models show good results with databases other than the one used for training. In this work two approaches were evaluated with three public databases, with results that indicate low generalization capacity of the models.*

Resumo. *Ataque de replay é uma falsificação de fala utilizada na tentativa de autenticação de locutor. Redes neurais profundas têm sido propostas como métodos para detecção de áudios fraudulentos. Tendo em vista a utilização desses modelos em aplicações reais, além de bom desempenho na aprendizagem, espera-se que o modelo obtido apresente bons resultados com bases de dados distintas da utilizada no treinamento. Neste trabalho, duas abordagens foram avaliadas com três bases de dados públicas, com resultados que indicam baixa capacidade de generalização dos modelos.*

1. Introdução

Sistemas de biometria de voz estão sendo amplamente utilizados nos mais diversos setores, como indústria automotiva, financeiro, saúde e educação [Khan et al. 2023]. Nesses casos, a autenticação do usuário é realizada por sistemas de verificação automática de locutor (*Automated Speaker Verification* – ASV), suscetíveis a ataques de falsificação (*spoofing*).

Ataque de *replay* consiste na apresentação a um ASV da reprodução de um áudio previamente gravado, com o objetivo de validar a fala do locutor como genuína. Esse tipo de falsificação ocorre de forma passiva e é difícil de ser detectado, uma vez que o sinal reproduzido apresenta semelhanças físicas (frequências, espectros, formas de ondas) ao original [Khan et al. 2023]. Para aumentar a confiabilidade, os sistemas ASV são combinados com sistemas que identificam a fala falsificada, também chamados de *antispoofing* [Alzantot et al. 2019].

A série de competições *Automatic Speaker Verification Spoofing and Countermeasures Challenge* (ASVspoof) promove, desde 2015, o desenvolvimento de métodos para

detecção de falsificação. A cada edição, uma base de dados é disponibilizada para ser utilizada no treinamento e na validação de contramedidas aos ataques [Alzantot et al. 2019, Lee et al. 2022, Nautsch et al. 2021].

De forma geral, os métodos iniciam com a extração de atributos (*features*), predominantemente baseados na análise espectral do áudio, como espectrogramas, cepstrogramas, coeficientes em escala mel e variações da análise de Fourier [Khan et al. 2023]. O processo de aprendizado ocorre utilizando-se esses atributos como entradas para um classificador, tipicamente uma rede neural. Frequentemente tem sido utilizadas redes convolucionais [Chettri et al. 2018, Korshunov et al. 2018] (*Convolutional Neural Network* - CNN) e suas variações, como a rede convolucional leve (*Light Convolutional Neural Network* - LCNN) [Lavrentyeva et al. 2017, Lavrentyeva et al. 2019] e a rede convolucional residual (*Residual Neural Network* - ResNet) [Alzantot et al. 2019, Zhang et al. 2021].

Usualmente, para utilização em uma aplicação real, é esperado que o modelo apresente uma boa capacidade de generalização [Korshunov and Marcel 2016], isto é, que seu desempenho não seja muito diferente quando comparados os resultados obtidos em dados distintos dos usados no treinamento.

O objetivo deste trabalho é estudar o desempenho de abordagens propostas à detecção de ataque de *replay* entre bases de dados distintas da utilizada no aprendizado das redes, de modo a contribuir para a discussão sobre generalização dos modelos.

2. Metodologia

Neste trabalho, duas abordagens de classificação foram avaliadas com três bases de dados públicas. O desempenho dos métodos foi mensurado pelo EER (*Equal Error Rate*), ponto de operação em que a taxa de falsa aceitação (*False Acceptance Rate* - FAR) e a taxa de falsa rejeição (*False Rejection Rate* - FRR) são iguais [Jain et al. 2008].

2.1. Bases de Dados

O estudo foi realizado com três bases de dados públicas: ASVspoof 2019¹, ASVspoof 2021² e REMASC³. As três bases foram gravadas em inglês e, portanto, a influência da variação de idioma não pode ser explorada nesse caso. O treinamento de todos os modelos neste trabalho utilizou o conjunto de treinamento da base de dados ASVspoof 2019.

As bases ASVspoof 2019 e ASVspoof 2021 são compostas por arquivos de áudio do tipo *flac*, com um canal e taxa de amostragem de 16kHz. Ambas possuem outros tipos de ataque além do ataque de *replay*, mas, para os experimentos deste trabalho, foram utilizados apenas os conjuntos relativos ao ataque denominado de acesso físico (*Physical Access* - PA), pois são os dados com o ataque de *replay*. Esses conjuntos são formados por 218.430 e 943.110 amostras, respectivamente.

A base de dados REMASC foi concebida visando sistemas controlados por voz (*Voice Controlled Systems* - VCS), em que a coleta do áudio ocorre a uma distância maior do locutor. Seu conjunto abrange 54.712 amostras, armazenadas em arquivos do tipo *wav*, multicanais, amostrados a 16kHz e 44kHz [Gong et al. 2019]. Os áudios foram

¹<https://datashare.ed.ac.uk/handle/10283/3336>

²<https://www.asvspoof.org/index2021.html>

³github.com/ndmobilecomplab/replay-attack

padronizados em monocanais a 16kHz, aplicando a média dos canais e redução da taxa de amostragem (*downsampling*) quando necessário.

As três bases de dados disponibilizam arquivos de protocolo, indicando quais dados devem ser usados para treinamento e teste, bem como a identificação do locutor e do áudio e sua classificação original (genuíno ou falso). Além disso, fornecem metadados como tamanho do ambiente e características dos dispositivos utilizados para coleta.

2.2. Modelos

Neste trabalho, foram avaliadas dois tipos de arquitetura de redes neurais, ResNet e LCNN. A ResNet avaliada usa como atributos de entrada a magnitude do espectro em escala logarítmica (*Log-magnitude STFT - Short-Time Fourier Transform*). Foi utilizado um modelo pré-treinado disponibilizado publicamente ⁴.

Quanto à rede LCNN, foi utilizada uma implementação disponibilizada publicamente ⁵ e usada em um estudo que avaliou diversos atributos como entrada para a rede [Lee et al. 2022, Lee 2024]. Neste caso, como não há modelo pré-treinado disponível, o treinamento foi executado utilizando os seguintes atributos: análise discreta arbitrária de Fourier (*arbitrary discrete Fourier analysis - ADFA*), cepstrogramas (CEPS e CEPS1724), *constant Q analysis (CQA)*, transformada discreta de cosseno (*discrete cosine transform - DCT*), análise discreta de Fourier em escala Mel (*Mel-scale discrete Fourier analysis - MDFA*) e espectrogramas (Spec e Spec1724). Os espectrogramas e cepstrogramas foram extraídos pela transformada rápida de Fourier (*Fast Fourier Transform - FFT*) utilizando uma janela de Blackman com comprimento 1024 (Spec e Ceps) e 1724 (Spec1724 e Ceps1724).

Ainda, como referência para comparação, foi usada a abordagem LFCC-LCNN disponibilizada como *baseline* para o desafio ASVspoof 2021 [Liu et al. 2023], que inclui um modelo pré-treinado.

3. Resultados e Discussão

Os treinamentos da abordagem RD-LCNN foram realizados ao longo de cem épocas e o foi escolhido o modelo obtido a partir da época com menor valor EER no conjunto de desenvolvimento da base ASVspoof 2019.

A Tabela 1 apresenta os resultados de todos os modelos avaliados nas três bases de dados, ASVspoof 2019, ASVspoof 2021 e REMASC. As colunas (a), mostram como referência os resultados relatados na literatura para os subconjuntos de desenvolvimento (Dev) e de avaliação (Eval) da base ASVspoof 2019 [Nautsch et al. 2021] e, nas demais colunas, os resultados obtidos nos experimentos deste trabalho.

O modelo pré-treinado da *baseline* LFCC-LCNN apresentou bom desempenho com o subconjunto *eval* da base de dados ASVspoof 2019, com EER = 2,43%, fato esperado, uma vez que o treinamento ocorreu com o subconjunto *train* dessa mesma base de dados. É importante observar que a mesma superou as *baselines* propostas em 2019, CQCC-GMM e LFCC-GMM, que apresentaram EER = 11,04% e EER = 13,54%, respectivamente [Nautsch et al. 2021]. Já com a base ASVspoof 2021, o desempenho (EER

⁴<https://github.com/nesl/asvspoof2019>

⁵<https://github.com/shihkuanglee/RD-LCNN/tree/main>

Tabela 1. Resultados do ERR (%) obtidos (ASVspooF 2019 (b), ASVspooF 2021 e REMASC) e reportados pela literatura (ASVspooF 2019 (a))

Modelo	ASVspooF 2019 (a)		ASVspooF 2019 (b)			ASVspooF 2021	REMASC
	Dev	Eval	Train	Dev	Eval	Eval	-
Baseline LFCC-LCNN	42,16	*	**	42,16	2,43	45,67	***
Spec - ResNet	3,85	3,81	**	3,85	7,07	43,21	48,49
ADFA - RD-LCNN	0,22	0,85	0,71	0,83	1,67	40,85	60,87
Ceps - RD-LCNN	0,13	0,37	0,20	0,17	0,41	37,13	49,21
Ceps1724 - RD-LCNN	0,28	0,71	0,23	0,28	0,80	49,77	51,66
CQA - RD-LCNN	0,35	0,74	0,75	0,76	0,97	46,72	49,78
DCT - RD-LCNN	1,44	2,90	1,58	0,69	11,66	38,15	56,85
MDFA - RD-LCNN	0,17	0,61	0,53	0,59	1,39	43,02	59,58
Spec - RD-LCNN	0,56	1,72	0,67	1,39	1,86	43,51	52,01
Spec1724 - RD-LCNN	0,20	0,92	0,58	0,79	1,65	35,65	50,54

* Resultado não disponível na literatura.

** Emprego de modelo pré-treinado; valor EER no conjunto train não reportado na literatura.

*** Processamento não realizado.

= 45,67%) foi similar ao relatado pela literatura (EER = 44,77%) [Liu et al. 2023] e, portanto, muito pior.

A abordagem ResNet para a base de dados ASVspooF 2019 do subconjunto *eval* apresentou resultado (EER = 7,07%) próximo ao da literatura (EER = 3,81%) e, embora 86% maior, ainda foi melhor que os das *baselines* do desafio de 2019: EER = 11,04% (B01 - CQCC-GMM) e EER = 13,54% (B02 - LFCC-GMM) [Nautsch et al. 2021]. Para as bases ASVspooF 2021 e REMASC observa-se um baixo desempenho, com valores EER acima de 40%.

Os modelos treinados da abordagem RD-LCNN apresentaram resultados para o subconjunto *dev* do ASVspooF 2019 próximos aos relatados na literatura. Para o subconjunto *eval* da base ASVspooF 2019, o atributo “DCT” expressou a maior diferença. A inferência nas bases de dados ASVspooF 2021 e REMASC também resultaram em valores EER muito piores, maiores que 35%.

4. Conclusão

Os resultados obtidos para a base de dados ASVspooF 2019 demonstram bom desempenho da *baseline* e das abordagens experimentadas, semelhantes aos relatados pela literatura. Observa-se que a *baseline* obteve ERR = 2,43% no ASVspooF 2019 - *eval*, e, embora não se tenha encontrado valor na literatura para efeito de comparação, esse resultado foi melhor que os de ambas as *baselines* do desafio de 2019.

Os altos valores EER obtidos com as bases de dados ASVspooF 2021 e REMASC ratificam a situação exposta por [Korshunov and Marcel 2016], apontando para uma baixa capacidade de generalização de todas as abordagens processadas.

Agradecimentos

Este projeto foi apoiado pelo Ministério da Ciência, Tecnologia e Inovações, com recursos da Lei no 8.248, de 23 de outubro de 1991, no âmbito do PPI-SOFTEX, coordenado pela Softex e publicado PDI 03, DOU 01245.023862/2022-14.

Referências

- Alzantot, M., Wang, Z., and Srivastava, M. B. (2019). Deep residual neural networks for audio spoofing detection. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2019-September:1078–1082.
- Chettri, B., Mishra, S., Sturm, B. L., and Benetos, E. (2018). A study on convolutional neural network based end-to-end replay anti-spoofing. *arXiv*.
- Gong, Y., Yang, J., Huber, J., MacKnight, M., and Poellabauer, C. (2019). Remasc: Realistic replay attack corpus for voice controlled systems. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, volume 2019-September, pages 2355–2359. International Speech Communication Association.
- Jain, A. K., Flynn, P., and Ross, A. A. (2008). *Handbook of Biometrics*. Springer.
- Khan, A., Malik, K. M., Ryan, J., and Saravanan, M. (2023). Battling voice spoofing: a review, comparative analysis, and generalizability evaluation of state-of-the-art voice spoofing counter measures. *Artificial Intelligence Review*, 56:513–566. 01.
- Korshunov, P., Gonçalves, A. R., Violato, R. P. V., Simões, F. O., and Marcel, S. (2018). On the use of convolutional neural networks for speech presentation attack detection. In IEEE, editor, *2018 IEEE 4th international conference on identity, security, and behavior analysis (ISBA)*, pages 1–8.
- Korshunov, P. and Marcel, S. (2016). Cross-database evaluation of audio-based spoofing detection systems. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, volume 08-12-September-2016, pages 1705–1709. International Speech and Communication Association.
- Lavrentyeva, G., Novoselov, S., Malykh, E., Kozlov, A., Kudashev, O., and Shchemelinin, V. (2017). Audio replay attack detection with deep learning frameworks. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, volume 2017-August, pages 82–86. International Speech Communication Association.
- Lavrentyeva, G., Novoselov, S., Tseren, A., Volkova, M., Gorlanov, A., and Kozlov, A. (2019). Stc antispoofing systems for the asvspoof2019 challenge. *arXiv*.
- Lee, S.-K. (2024). Arbitrary discrete fourier analysis and its application in replayed speech detection. *arXiv*.
- Lee, S.-K., Tsao, Y., and Wang, H.-M. (2022). Detecting replay attacks using single-channel audio: The temporal autocorrelation of speech. In *Proceedings of 2022 APSIPA Annual Summit and Conference*. 2022 APSIPA Annual Summit and Conference.
- Liu, X., Wang, X., Sahidullah, M., Patino, J., Delgado, H., Kinnunen, T., Todisco, M., Yamagishi, J., Evans, N., Nautsch, A., and Lee, K. A. (2023). Asvspoof 2021: Towards spoofed and deepfake speech detection in the wild. *IEEE/ACM Transactions on Audio Speech and Language Processing*, 31:2507–2522.
- Nautsch, A., Wang, X., Evans, N., Kinnunen, T., Vestman, V., Todisco, M., Delgado, H., Sahidullah, M., Yamagishi, J., and Lee, K. A. (2021). Asvspoof 2019: spoofing countermeasures for the detection of synthesized, converted and replayed speech. *arXiv*.

Zhang, Z., Yi, X., and Zhao, X. (2021). Fake speech detection using residual network with transformer encoder. In *IH and MMSec 2021 - Proceedings of the 2021 ACM Workshop on Information Hiding and Multimedia Security*, pages 13–22. Association for Computing Machinery, Inc.

Avaliação de arquiteturas de síntese de fala generativa com abordagens de espectrograma e fim-a-fim em cenários *low-resource* para clonagem de voz

Bruno C. dos S. Ribeiro¹, Gustavo H. dos S. Figueiredo¹,
Leonardo H. da S. Correia¹, Mário Uliani Neto¹, Fernando O. Runstein¹,
Ricardo P. V. Violato¹, Marcus Lima²,

¹CPQD - Centro de Pesquisa e Desenvolvimento, Campinas, SP, Brasil

²Pontifícia Universidade Católica de Campinas, SP, Brasil

Resumo. *O artigo compara modelos de síntese de fala com arquiteturas baseadas em espectrograma e fim-a-fim, com o objetivo de determinar a capacidade de clonagem de voz em cenário low-resource. Foram avaliados conjuntos de treinamento de adaptação com diferentes quantidades de fala para clonagem de uma voz alvo, e o tempo necessário para realizar o treinamento. O modelo VITS mostrou-se mais eficiente, alcançando os melhores resultados no teste de qualidade perceptual no cenário low-resource com dados no idioma português, e completou o treinamento em menos tempo, quando comparado com o Tacotron2.*

1. Introdução

A síntese de fala tem sido um campo de intenso estudo e inovação ao longo dos últimos anos, com avanços significativos impulsionados pelos rápidos progressos na área de inteligência artificial generativa. Dentro deste contexto, diversas abordagens têm sido exploradas, incluindo as arquiteturas baseadas em espectrogramas e as abordagens fim-a-fim.

As arquiteturas Tacotron [Wang et al. 2017] e Tacotron2 [Shen et al. 2018] têm sido amplamente estudadas e aplicadas, demonstrando a capacidade de converter texto em fala natural por meio da geração de espectrogramas intermediários, que são posteriormente transformados em sinais de fala através de vocoders, como as arquiteturas WaveNet [van den Oord et al. 2016] e HiFi-GAN [Kong et al. 2020]. Apesar dos resultados promissores, esses modelos frequentemente requerem grandes quantidades de dados e longos períodos de treinamento para atingir um nível satisfatório de qualidade e naturalidade na fala.

As abordagens mais recentes de síntese de fala, como o modelo VITS (do inglês, *Variational Inference Text-to-Speech*) [Kim et al. 2021], propõem uma estratégia fim-a-fim que elimina a necessidade de um estágio intermediário explícito de geração do espectrograma, combinando de forma eficaz a geração e a codificação do sinal de fala em um único fluxo de trabalho. Este método tem mostrado potencial em reduzir significativamente a quantidade de dados necessários para o treinamento, bem como o tempo total para alcançar resultados de alta qualidade.

A eficiência da síntese de fala em cenários com recursos limitados (*low-resource*) é uma área de interesse crescente, especialmente para idiomas com menor disponibilidade de dados anotados, como o caso do português. Trabalhos recentes têm investigado a

eficácia de diferentes modelos em condições *low-resource*, abordando desafios específicos como a qualidade da fala sintetizada, a adaptabilidade de modelos pré-treinados para novos falantes e a eficiência computacional do processo de treinamento [Lux et al. 2022].

O objetivo deste artigo é comparar as arquiteturas baseadas em espectrogramas e fim-a-fim no contexto de clonagem de voz em português, com ênfase no desempenho do VITS versus o Tacotron2. O objetivo é comparar os modelos em cenários *low-resource* e quantificar o número mínimo de dados e tempo de treinamento necessários para atingir resultados de alta qualidade. Os resultados baseiam-se em métricas de qualidade objetiva e subjetiva, e na análise do tempo de treinamento. Esperamos fornecer *insights* práticos para a escolha e implementação de modelos de síntese de fala com voz personalizada em condições de dados restritos, contribuindo para a eficiência e a acessibilidade da tecnologia de síntese de fala em uma ampla gama de aplicações para o idioma português do Brasil.

2. Metodologia

O treinamento dos modelos foi realizado utilizando duas bases de fala no idioma Português Brasileiro: (i) o *TTS-Portuguese Corpus* [Casanova et al. 2022], composto por textos de domínio público provenientes tanto da Wikipédia quanto do Chatterbot-corpus (um corpus criado originalmente para a construção de *chatbots*), contendo aproximadamente 10 horas e 28 minutos de fala de um único locutor masculino, gravada com taxa de amostragem de 48 kHz e 16 bits, tendo 3.632 áudios no formato WAV linear, com um range de duração de 0,67 a 50,08 segundos (todos os cliques de áudio com duração superior a 20 segundos foram removidos do treinamento); (ii) uma base de fala proprietária do CPQD composta por um locutor masculino contendo 20 minutos de fala, gravada com taxa de amostragem de 48kHz, 16 bits e formato PCM linear, contendo os arquivos de áudio e as transcrições ortográficas correspondentes.

O treinamento foi realizado a partir do repositório do VITS¹, que foi adaptado para a inclusão de fonemas do idioma português do Brasil, realizado através do uso do módulo *Phonemizer*² em conjunto com a *pipeline* de preparação de dados.

O treinamento dos modelos base ocorreram ao longo de 80 horas e 2.000 épocas no dataset *TTS-Portuguese Corpus*. A partir do último *checkpoint* gerado pelo modelo base, foram realizados *fine-tunings* trocando os dados de treinamento pela base proprietária com a voz do locutor masculino, usando conjuntos de treinamento com 20, 15, 10 e 5 minutos de fala visando avaliar a quantidade mínima de dados necessários para obter síntese de boa qualidade. O objetivo do *fine-tuning* é adaptar o modelo base para as características da voz alvo, ou seja, realizar a clonagem de voz. Após apenas 1 hora de treinamento de *fine-tuning* usando 20 minutos de fala, foram observados resultados de alta qualidade tanto no VITS como no Tacotron2. A qualidade melhorou ainda mais após 20 horas de treinamento. Ambos utilizaram o vocoder HiFi-GAN, sendo que no caso do Tacotron2 o vocoder foi treinado de forma independente. Para os conjuntos de treinamento menores, a seção 3 apresenta os resultados obtidos.

¹<https://github.com/jaywalnut310/vits/>

²<https://pypi.org/project/phonemizer/3.0.1/>

3. Resultados

Para avaliar a qualidade da fala sintetizada resultante foram utilizadas medidas objetivas e subjetivas. As métricas objetivas foram o MCD (do inglês, *Mel-Cepstral Distortion*) e o F0 RMSE (do inglês, *Log-F0 Root Mean Square Error*) [Hayashi et al. 2021]. Para a avaliação subjetiva foi utilizada a métrica MOS (*Mean Opinion Score*), em um experimento que contou com 15 avaliadores não especialistas.

A métrica MCD, calculada por meio do repositório *TTS Objective Metrics*³, quantifica a distância entre dois sinais de fala. Quanto menor o valor MCD, mais semelhantes são as vozes. A qualidade da voz sintetizada foi avaliada com base no conjunto de teste, com frases separadas para validação. Ao comparar a voz sintetizada resultante do modelo de *fine-tuning* obtido com 20 minutos, com a voz original gravada, obteve-se valores de MCD entre 1.6 e 1.78. A métrica MCD mostra valores próximos de 0, indicando que o modelo é capaz de gerar fala sintetizada próxima da fala gravada. Para o F0 RMSE, aplicada nas mesmas sentenças, foram obtidos valores entre 0.18 e 0.34. Os resultados reforçam a alta qualidade da fala sintetizada.

3.1. Avaliação Subjetiva

Para a avaliação subjetiva foi utilizado o servidor webMUSHRA⁴. Um grupo de 15 avaliadores não especialistas ouviram um conjunto de amostras e atribuíram notas de 0 a 100 com base na naturalidade da voz, sendo 0 nada natural e 100 muito natural. Esse processo permitiu realizar uma análise subjetiva da qualidade do áudio sintetizado, proporcionando uma análise mais fidedigna da percepção humana em relação ao desempenho dos modelos. As avaliações mostram uma melhor qualidade do VITS em relação ao Tacotron2. A Figura 1 mostra o *boxplot* com os dados do teste subjetivo, utilizando áudios sintetizados por modelos obtidos através do *fine-tuning* com diferentes conjuntos de treinamento da voz alvo. Na legenda, 400 representa o conjunto com 20 minutos de fala, 300 indica 15 minutos, 200 indica 10 minutos e 100 indica o conjunto com 5 minutos de fala.

Os resultados indicam que o VITS (C1) consistentemente recebeu avaliações mais altas em comparação ao Tacotron2 (C2). O desvio padrão menor do VITS em comparação ao Tacotron2 em todos os conjuntos de treinamento indica que as opiniões dos usuários sobre a qualidade do áudio gerado pelo VITS são mais consistentes e robustas.

No teste realizado com o conjunto contendo 5 minutos de fala de treinamento, o VITS teve uma média de 71,49 enquanto o Tacotron2 teve 67,49. Essa diferença foi consistente em todos os conjuntos de treinamento (20, 15, 10 e 5 minutos). No entanto, a diferença aumenta com um volume maior de dados, sugerindo que o VITS não apenas produz áudio de melhor qualidade, mas também que melhora mais conforme a quantidade de dados de treinamento aumenta.

4. Conclusão

O objetivo principal deste trabalho foi comparar as arquiteturas de síntese de fala generativa com abordagens de espectrograma (Tacotron2) e fim-a-fim (VITS) em cenários

³<https://github.com/AI-Unicamp/TTS-Objective-Metrics>

⁴<https://github.com/audiolabs/webMUSHRA/>

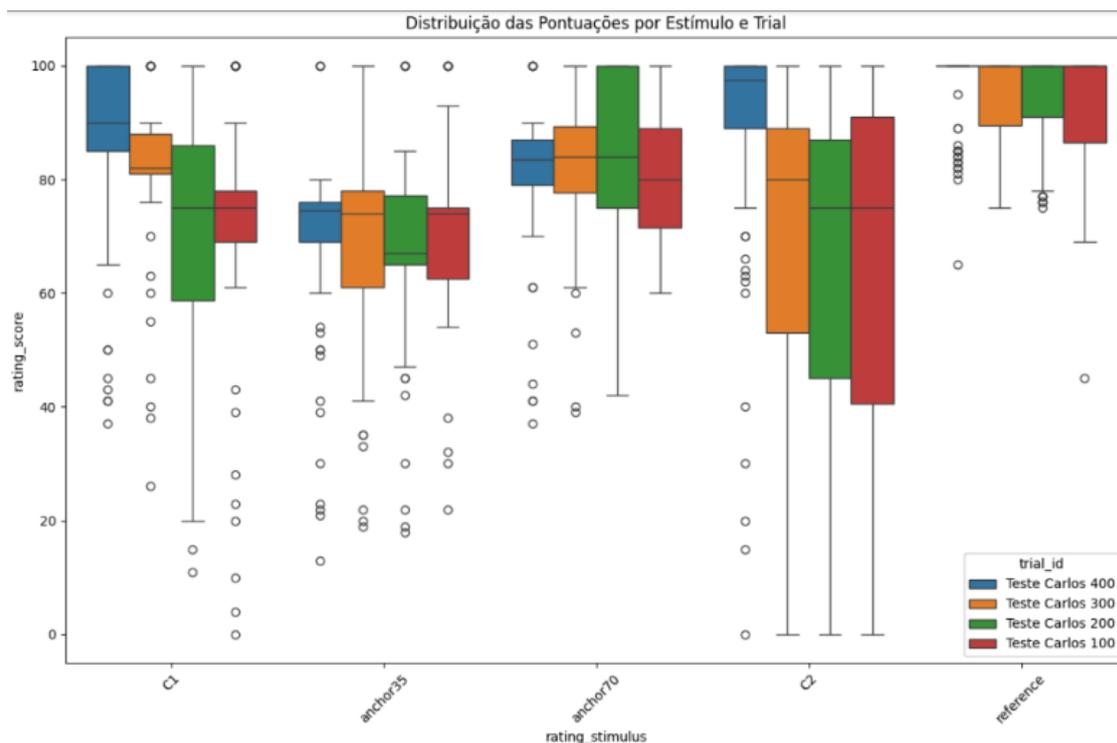


Figura 1. Boxplot da distribuição das pontuações por estímulo. C1 representa o VITS, e C2 representa o Tacotron2.

low-resource, com uso de até 5 minutos de fala no treinamento de *fine-tuning*, para clonagem de voz; ou seja, avaliar a capacidade de adaptação dos modelos base pré-treinados fazendo uso de dados limitados de uma nova voz personalizada.

O modelo VITS, quando treinado com 20 minutos, mostrou resultados com alta qualidade após apenas 1 hora de treinamento. Por outro lado, o Tacotron2, sob as mesmas condições, apresentou maior variabilidade e menor consistência na qualidade do áudio sintetizado. Mesmo quando treinado com 5 minutos o VITS apresentou boa qualidade e baixa variância. Ao comparar o tempo de treinamento, o modelo VITS mostrou-se mais eficiente, alcançando bons resultados em menos tempo e com menos dados em relação ao Tacotron2.

Os resultados indicam que o VITS não só oferece uma síntese de fala de melhor qualidade, com maior similaridade à voz original e menor variância entre as amostras sintetizadas, mas também é mais eficiente em termos de tempo de treinamento em cenários *low-resource*.

5. Agradecimentos

Este projeto foi apoiado pelo Ministério da Ciência, Tecnologia e Inovações, com recursos da Lei no 8.248, de 23 de outubro de 1991, no âmbito do PPI-SOFTEX, coordenado pela Softex e publicado PDI 03, DOU 01245.023862/2022-14.

Referências

- Casanova, E., Junior, A. C., Shulby, C., Oliveira, F. S. d., Teixeira, J. P., Ponti, M. A., and Aluísio, S. (2022). Tts-portuguese corpus: a corpus for speech synthesis in brazilian portuguese. *Language Resources and Evaluation*, 56(3):1043–1055.
- Hayashi, T., Yamamoto, R., Yoshimura, T., Wu, P., Shi, J., Saeki, T., Ju, Y., Yasuda, Y., Takamichi, S., and Watanabe, S. (2021). Espnet2-tts: Extending the edge of tts research.
- Kim, J., Kong, J., and Son, J. (2021). Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech.
- Kong, J., Kim, J., and Bae, J. (2020). Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis.
- Lux, F., Koch, J., and Vu, N. T. (2022). Low-resource multilingual and zero-shot multipeaker tts.
- Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z., Chen, Z., Zhang, Y., Wang, Y., Skerry-Ryan, R., Saurous, R. A., Agiomyrgiannakis, Y., and Wu, Y. (2018). Natural tts synthesis by conditioning wavenet on mel spectrogram predictions.
- van den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., and Kavukcuoglu, K. (2016). Wavenet: A generative model for raw audio.
- Wang, Y., Skerry-Ryan, R., Stanton, D., Wu, Y., Weiss, R. J., Jaitly, N., Yang, Z., Xiao, Y., Chen, Z., Bengio, S., Le, Q., Agiomyrgiannakis, Y., Clark, R., and Saurous, R. A. (2017). Tacotron: Towards end-to-end speech synthesis.

Leveraging Structured Data Input for Effective Chatbot Integration in Enterprises

Caio Siqueira¹, Orlando Guilarte¹, Giuliano Ferreira¹, Omar Leiva¹

¹Diretoria de Sistemas de Informação - PUC-Rio
Rua Marquês de São Vicente 225 – 22451-900 - Gávea - Rio de Janeiro, Brasil

{csiqueira, ofonsek, giuliano.biblioteca}@puc-rio.br

omarleivac@aluno.puc-rio.br

Abstract. *This paper introduces an approach for integrating structured data into chatbot applications. Utilizing our Mindmap tool, which hierarchically organizes data and maps nodes to actions, we developed an augmented JSON schema to improve chatbot contextual understanding and response accuracy. By applying the Langchain suite and Retrieval-Augmented Generation techniques, our method enhances data retrieval and processing from a vector store, significantly improving interaction relevance.*

Keywords: *Chatbot Integration; Structured Data; RAG; Langchain*

1. Introduction

In recent years, the evolution of consumer-facing software has led to heightened expectations among corporate users for more intuitive and natural interactions with computer systems. Traditionally, user interfaces in corporate environments have relied on window-based interaction paradigms. However, the advent of sophisticated natural language processing (NLP) technologies has begun to shift this paradigm, making natural language interfaces increasingly desirable for enterprise applications [Weiyang et al. 2019]. A significant milestone in this transition was the public release of ChatGPT by OpenAI [OpenAI 2023]. The underlying technology behind ChatGPT is based on a Large Language Model (LLM), a type of machine learning model trained on extensive datasets containing diverse textual content. The scale of these training datasets allows LLMs to effectively respond to a wide range of user inputs. However, deploying LLMs in corporate environments presents unique challenges. A key limitation is that these models are typically not trained on proprietary data from private institutions, which must remain confidential to protect organizational integrity and data privacy. To address this issue, researchers have explored the use of knowledge graphs to further enhance LLMs [Wen et al. 2023], and techniques such as Retrieval-Augmented Generation (RAG) have been developed to provide LLMs with context derived from private datasets, thereby enhancing their relevance and accuracy in enterprise settings [Lewis et al. 2020]. Despite the effectiveness of these techniques, they often fall short of meeting the demands of corporate software development, where frequent updates and rapid responses to organizational changes are common. These approaches also face challenges such as difficulty in incorporating new knowledge and explaining their reasoning processes [Wen et al. 2023].

This study proposes a mechanism for structuring data more effectively to feed LLM-based chatbots with the contextual information necessary to provide accurate and

contextually relevant responses within corporate environments. Given the stringent requirements of corporate settings, where control over generated outputs is paramount, the approach outlined in this work prioritizes supervised interaction models over autonomous agent-based systems. Unlike general public applications, where disclaimers can mitigate the risks of inaccurate or flawed outputs, corporate environments require a higher level of oversight to prevent potential adverse outcomes.

2. Conception

With the growing demand from our key clients, particularly those forming the consumer base of our business model, for a more integrated and sophisticated LLM chatbot interface, we developed an approach that leverages existing structured chatbot systems and the data already mapped within these systems.

The structured chatbot system organized data into a hierarchical, tree-like structure, where each parent node represented a topic, and the child nodes indicated possible responses or subtopics associated with that topic. This setup ensured a well-defined and navigable dialogue structure. With the goal of creating an LLM-based chatbot integrated with RAG — which employs a vector store for semantic retrieval — we developed a tool capable of exporting this structured data for integration into the new system.

However, we identified significant limitations in our current software, which only supported data generation in RTF (Rich Text Format), a format unsuitable for our needs. To overcome this challenge, we developed a new tool to manage the registration of nodes and their associated child nodes. This led to the creation of a web-based interface (Mindmap), built using advanced JavaScript frameworks (React with Next.js [Vercel, Inc. 2024]).

The introduction of this tool significantly expanded the scope of the project, allowing us to move beyond simple topic and child text nodes, enabling the creation of more complex data structures tailored to specific objectives. One of the critical features developed was the mapping of nodes to actions, which dictate the operations to be executed by the chatbot. The text registered within each node is subsequently parsed as parameters for the corresponding action.

With the development of the Mindmap tool, we now have a robust platform that allows for the systematic registration of necessary data and facilitates its export into formats more suitable for subsequent processing.



(a) Screenshot of the Mindmap tool

(b) Snippet of the augmented json

Figure 1. Generated augmented json

3. Proposed pipeline

The Mindmap tool exports hierarchical data as a structured JSON file, which is processed into an augmented JSON using Python scripts within the Langchain suite [LangChain Documentation], currently associated with the GPT-4 model from OpenAI. This augmented file incorporates questions and answers, derived from the hierarchy of each node, its actions, and related metadata.

These questions and answers are generated based on the node's actions and predefined profiles. The resulting JSON is then used to populate a vector store database, serving as the foundation for the chatbot's RAG functionalities. The design is agnostic concerning the chatbot engine's method of consuming this data to construct its vector store or run the inference itself. For example, a collaborating research group utilizes the vector store without incorporating questions and answers, using them later to validate the generated dataset. In contrast, internally, we employ the questions and answers to create documents directly in the vector store. A comparative study on the efficacy of these strategies for various use cases could be the focus of future research.

Table 1. Sample of node actions

Action name	Action description
Fetch data on a website	The node contains a URL and a CSS selector to fetch data. Additionally, it contains a JUDGE text to check if the data is similar to what is expected.
Use a vector store for document	Given a document link, build a vector store specific to that document. Once an initial query matches that the answer should come from the document, a second LLM query is used on that specific vector store.
Fetch an API	Translate the user input into an API call. Translate the return into readable output to serve for the user. The node has instructions on how this should be done.
Serve node text	Use the node text as context to provide an answer to the user input.

4. Data structure

The output augmented JSON schema encapsulates several key elements, including the list of actions within the exported object, the profiles used for generating questions and answers, and the hierarchical structure of nodes and their children. Each node is assigned a unique identifier and an update timestamp, generated by our Mindmap application, to facilitate efficient updates in downstream consumer applications.

To enhance the functionality of each node, we introduced a supplementary structure termed "helper." Each action within a node is associated with a helper, which consists of parameters parsed during the generation of the augmented JSON. Internally, our team loads these helpers into the vector database, enabling their retrieval at runtime via Langchain and Python.

Looking forward, a significant improvement on our roadmap involves integrating the Mindmap frontend tool with the Python backend responsible for generating the

augmented JSON structure. We also plan to migrate the generated content, including questions and answers, into our relational database alongside the nodes. This integration will streamline the workflow, allowing for direct export of the augmented JSON from the Mindmap application, thereby enhancing the user experience and operational efficiency.

Table 2. Sample profiles used when generating questions and answers

Question Profile	Profile description	Used in
Computer science student	You are a computer science student who focuses your input using direct messages	Questions
Language student	You are a language student with rich vocabulary	Questions
Internet user	You are an unknown internet user with poor grammar which basically uses keywords when interacting with systems	Questions
Institutional chatbot	You are an organization chatbot, which needs to answer in a formal way never betraying the ideals of the organization	Answers

5. Conclusion

The work we present proposes an innovative and cohesive approach to integrating the entire workflow in creating a chatbot application that is closely coupled with its underlying data. By leveraging structured data input and the advanced functionalities of our Mindmap tool, we have established a solid foundation that not only supports the continuous generation of enhanced JSON structures but also facilitates real-time data retrieval and processing through advanced technologies such as Langchain and Python.

The management of the Mindmap software is designed to be in the hands of the process owners, ensuring that those who understand the intricacies of each process are directly involved in its configuration and oversight. For example, one of our chatbot instances is currently managed by our Process Management Office (EP) ¹.

Our collaboration with the partner research group, the Applied Computational Intelligence Laboratory (ICA) ², has proven invaluable, providing critical insights and valuable feedback for the continuous improvement of our tool.

Looking ahead, we are confident that the continued development and refinement of this tool will further enhance our ability to integrate complex data structures with chatbot applications, ultimately contributing to more intelligent and responsive systems across the organization. This work not only demonstrates the potential of structured data integration but also lays the groundwork for future innovations in the field of enterprise chatbot solutions.

¹Escritório de Processos - <https://ep.dsi.puc-rio.br>

²Laboratório de Inteligência Computacional Aplicada - <https://ica.ele.puc-rio.br>

References

- LangChain Documentation. Langchain: Building applications with llms through composability. n.d.
- Lewis, P., Oguz, B., Rinott, R., Riedel, S., and Stenetorp, P. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS 2020)*, Vancouver, Canada.
- OpenAI (2023). Gpt-4 technical report. Technical report, OpenAI.
- Vercel, Inc. (2024). *React Foundations: About React and Next.js*. Next.js Documentation.
- Weiyang, K., Pham, D. N., Eftekharypour, Y., and Pheng, A. J. (2019). Benchmarking nlp toolkits for enterprise application. In *PRICAI 2019: Trends in Artificial Intelligence: 16th Pacific Rim International Conference on Artificial Intelligence, Cuvu, Yanuca Island, Fiji, August 26-30, 2019, Proceedings, Part III 16*, pages 289–294. Springer.
- Wen, Y., Wang, Z., and Sun, J. (2023). Mindmap: Knowledge graph prompting sparks graph of thoughts in large language models. *arXiv preprint arXiv:2308.09729*.

Anomaly Detection in Text Data: A Semi-Supervised Approach Applied to the Portuguese Domain

Fabio Masaracchia Maia¹, Anna Helena Reali Costa¹

¹Escola Politécnica, Universidade de São Paulo, São Paulo, Brazil

fabio.masaracchia@gmail.com, anna.reali@usp.br

Abstract. *Anomaly detection, driven by advancements in machine learning and deep learning, has gained significant importance across various fields. However, its application to unstructured textual data, particularly in Portuguese, remains underexplored. In textual analysis, these techniques are crucial for detecting deviations within text collections. This paper investigates state-of-the-art methods for anomaly detection in Portuguese text corpora and introduces a new, flexible loss function designed to enhance detection across different contamination levels. By evaluating these methods on benchmark datasets, specifically in the contexts of hate speech detection and sentiment analysis, we address existing challenges and contribute to the development of more effective anomaly detection techniques for Portuguese text data.*

1. Introduction

Anomaly detection refers to the identification of patterns in data that deviate from expected norms [Chandola et al. 2009]. Anomalies, often termed outliers or exceptions, are distinct from the majority of observations that define the “normal” pattern. While anomaly detection techniques have been extensively applied to structured data, such as continuous and categorical variables [Boutalbi et al. 2023], less attention has been given to unstructured data like text — the focus of this work.

Anomaly detection began with statistical methods in the late 19th century [Edgeworth 1887] and has since expanded, with deep learning broadening its scope to unstructured domains like images and text [Chandola et al. 2009, Pimentel et al. 2014]. However, its application to textual anomaly detection remains limited [Pang et al. 2019]. Detecting anomalies in text data is particularly challenging due to the variety of linguistic levels involved, such as spelling, syntax, and semantics [Xu et al. 2023b]. Leveraging deep learning’s ability to model complex patterns has led to significant advancements in the field.

Traditionally treated as an unsupervised task due to the absence of ground truth labels, anomaly detection has employed techniques like autoencoders and GANs [Pang et al. 2019]. Popular approaches in this domain include DeepSVDD [Ruff et al. 2018] and Deep Isolation Forest [Xu et al. 2023a], both focusing on modeling normal data and identifying deviations as anomalies. However, recent semi-supervised neural approaches, such as DevNet [Pang et al. 2019] and DeepSAD [Ruff et al. 2020], have shown improved detection accuracy by integrating limited labeled anomalies into the training process [Xu et al. 2023b], bridging the gap between unsupervised and supervised

learning. Despite recent advancements, there remains a significant gap in comprehensive research focused on anomaly detection in Portuguese text corpora.

In this paper, we extend neural network-based anomaly detection techniques to handle these complexities in Portuguese text data. Furthermore, we propose a change in the loss function in order to establish a compromise between the samples that correspond to anomalies in relation to the others. Experiments show that this approach is quite promising. To effectively address the unique challenges of representing textual data, we employ two pre-trained BERT-based models, checking the strengths and weaknesses of each representation in the different tasks.

2. Methodology

2.1. Problem Definition

Given a dataset $X = \{x_1, x_2, \dots, x_{N+K}\}$, where $U = \{x_1, x_2, \dots, x_N\}$ is unlabeled data and $K = \{x_{N+1}, \dots, x_{N+K}\}$ represents labeled anomalies ($K \ll N$), the goal is to train a model to identify these rare anomalies. This task is challenging due to the imbalance between the large unlabeled set and the small labeled anomaly set. The process involves two key steps:

1. **Embedding Transformation:** Data X is transformed into embeddings $Z = \{z_1, z_2, \dots, z_{N+K}\}$, with each z_i being a vector in \mathbb{R}^d .
2. **Scoring Function:** A neural network learns a scoring function $\phi : Z \rightarrow \mathbb{R}$ to ensure that $\phi(z_i) > \phi(z_j)$ when z_i is an anomaly and z_j is normal, minimizing the use of labeled examples.

We adopted the DevNet model due to its demonstrated good performance obtained in studies considering textual domain [Xu et al. 2023b] along with its ability to effectively manage high-dimensional spaces, such as embeddings. Additionally, the model's interpretable loss function, based on a straightforward Z-score strategy, provides valuable insights that can be later used to assess text identified as anomalies.

2.2. DevNet for Anomalous Text

The DevNet algorithm [Pang et al. 2019] introduces a semi-supervised approach that learns an interpretable outlier scoring function, $\phi(z; \Theta)$, using a Z-score deviation loss. While the original formulation is based on raw data points x , we denote the embeddings as z to reflect the transformed data representations. This approach assumes a prior normal distribution over anomaly scores, modeled with l random objects $r_i \in \mathbb{R}$ sampled from a standard normal distribution $\mathcal{N}(\mu_R, \sigma_R)$:

$$\text{dev}(z) = \frac{\phi(z; \Theta) - \mu_R}{\sigma_R}, \quad (1)$$

where μ_R and σ_R are the mean and standard deviation of anomaly scores within the reference distribution. This deviation is then incorporated into a contrastive loss function to enhance the distinction between anomalous and normal samples, where y indicates anomaly status, and a ensures a minimum separation between classes [Pang et al. 2019].

$$L(\phi(z; \Theta), \mu_R, \sigma_R) = (1 - y)|\text{dev}(z)| + y \cdot \max(0, a - \text{dev}(z)), \quad (2)$$

2.3. Proposal

To provide flexibility, the parameter $\eta \in [0, 1]$ is introduced in the DevNet loss function given in Eq. 2, controlling the balance between regular and anomalous samples,

$$L(\phi(x; \Theta), \mu_R, \sigma_R) = (1 - \eta)(1 - y)|\text{dev}(x)| + \eta \cdot y \cdot \max(0, a - \text{dev}(x)), \quad (3)$$

The η parameter adjusts the model’s emphasis on anomalies, allowing adaptation to varying levels of contamination (i.e., percentage of labeled anomalies) and data availability. We investigate different proportions of labeled anomalies to assess the robustness of the solution in various scenarios, aiming to determine the minimum amount of labeled data needed for good performance. Additionally, we employ two distinct text representation strategies: the monolingual BERTimbau model [Souza et al. 2020], specifically designed for processing Portuguese text, and the multilingual Sentence-BERT (SBERT) [Reimers and Gurevych 2019], which generates sentence-level embeddings across multiple languages, including Portuguese. Our customized DevNet implementation, named η -DevNet, was evaluated against its original version using both representation strategies.

3. Experiments and Results

3.1. Experiments

To evaluate the performance of different representation methods and loss functions, we first tested η values ranging from 0.5 to 1 using the BERTimbau embedding strategy, where $\eta = 0.5$ corresponds to the original DevNet formulation. The other parameters were adopted from the DevNet reference: $a = 5$, $l = 5000$, $\mu_R = 0$, and $\sigma_R = 1$. Contamination levels were adjusted by introducing between 5 and 1000 anomalies across the experiments, with anomalies randomly selected. After identifying the optimal η value, we applied it in subsequent experiments to compare both embedding strategies across different datasets. The mean ROC-AUC values were calculated over 10 experimental runs for each scenario.

3.2. Dataset

We evaluate our approach using two Brazilian datasets. The first, Told-Br [Leite et al. 2020], contains 21,000 labeled instances of tweets tagged with hate speech, categorized into themes such as homophobia, racism, and misogyny, with hate speech serving as the anomaly class. The second dataset, UTLC-Movies [Sousa et al. 2019], comprises over one million movie reviews. From this dataset, we sampled 40,000 reviews for sentiment analysis, where negative sentiment is treated as the anomaly class.

3.3. Results

Figure 1, shows performance and stability improvements as η is adjusted, with $\eta = 0.7$ yielding optimal performance. This value was subsequently used for further analysis. The results shown in Table 1 outline these results for both tasks, demonstrating that in most cases, the adapted loss function led to performance improvements. Reaching a reasonable level of accuracy requires a minimum threshold of labeled examples, which varies with task complexity. In our experiments, sentiment analysis needed only 0.87% of labeled anomalies to achieve a ROC-AUC of 0.85, while hate speech detection required 2.59% to

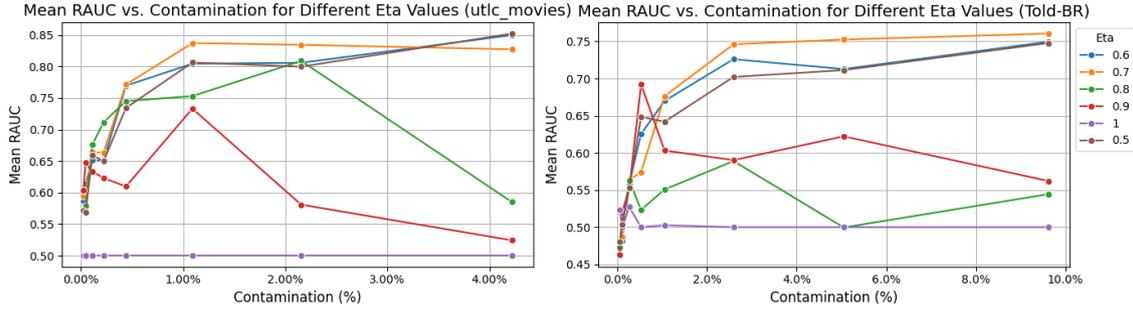


Figure 1. η Comparison across different η values with varying amounts of labeled anomalies using BERTimbau embedding strategy.

reach a ROC-AUC of 0.73. This discrepancy likely arises from the greater complexity of hate speech detection, which involves subtle linguistic nuances and diverse expressions. Additionally, pre-trained models may not fully capture slang and politically specific contexts, which are common in hate speech but may be underrepresented during training.

Table 1. Comparison of ROC-AUC values across different scenarios and contamination levels for UTLC-Movies and Told-BR datasets when $\eta = 0.7$, where % refers to the contamination level.

Nb. Outliers	UTLC-Movies					Told-BR				
	%	B_η	BD	M_η	MD	%	B_η	BD	M_η	MD
5	0.02	0.58	0.57	0.52	0.52	0.05	0.46	0.48	0.50	0.50
10	0.04	0.62	0.58	0.54	0.57	0.11	0.48	0.50	0.53	0.53
25	0.09	0.66	0.65	0.60	0.68	0.27	0.57	0.57	0.54	0.60
50	0.18	0.70	0.69	0.64	0.66	0.53	0.58	0.63	0.56	0.60
100	0.35	0.78	0.71	0.71	0.70	1.05	0.66	0.63	0.61	0.60
250	0.87	0.82	0.76	0.75	0.69	2.59	0.73	0.68	0.56	0.52
500	1.73	0.83	0.83	0.81	0.79	5.05	0.75	0.73	0.50	0.51
1000	3.41	0.85	0.84	0.81	0.75	9.62	0.76	0.76	0.50	0.52

Acronyms: BERTimbau η -loss (B_η), BERTimbau Devnet loss (BD), multilingual SBERT η -loss (M_η), multilingual SBERT Devnet loss (MD).

Our results show that the BERTimbau representation [Souza et al. 2020] consistently outperformed the multilingual model across tasks. This advantage can be traced to BERTimbau’s specialization in Portuguese, allowing it to capture more intricate linguistic nuances, such as idiomatic expressions and regional variations.

4. Conclusion and Future Work

This study shows that BERTimbau, tailored for Portuguese, consistently outperforms multilingual models in anomaly detection, with the customized loss function providing notable improvements. These results highlight the potential of semi-supervised methods for tasks like harmful content detection and sentiment analysis in Portuguese contexts with limited labeled data.

Future work may expand this approach to related tasks such as topic modeling, fake news detection, and fraud detection. Although some labeling effort is still required for good performance, the small amount of labeled data needed makes this approach feasible in resource-constrained scenarios. Furthermore, the promising advances in Large Language Models (LLMs) could not only serve as valuable tools for benchmarking but also automate anomaly tagging, reducing manual effort and enhancing adaptability and scalability across various real-world applications.

References

- Boutalbi, K., Loukil, F., Verjus, H., Telisson, D., and Salamatian, K. (2023). Machine learning for text anomaly detection: A systematic review. In *2023 IEEE 47th Annual Computers, Software, and Applications Conference (COMPSAC)*, pages 1319–1324.
- Chandola, V., Banerjee, A., and Kumar, V. (2009). Anomaly detection: A survey. *ACM Computing Surveys*, 41(3):71–97.
- Edgeworth, F. Y. (1887). Xli. on discordant observations. *Philosophical Magazine Series 1*, 23:364–375.
- Leite, J. A., Silva, D., Bontcheva, K., and Scarton, C. (2020). Toxic language detection in social media for Brazilian Portuguese: New dataset and multilingual analysis. In Wong, K.-F., Knight, K., and Wu, H., editors, *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 914–924, Suzhou, China. Association for Computational Linguistics.
- Pang, G., Shen, C., and van den Hengel, A. (2019). Deep anomaly detection with deviation networks. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD '19*, page 353–362, New York, NY, USA. Association for Computing Machinery.
- Pimentel, M. A., Clifton, D. A., Clifton, L., and Tarassenko, L. (2014). A review of novelty detection. *Signal Processing*, 99:215–249.
- Reimers, N. and Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In Inui, K., Jiang, J., Ng, V., and Wan, X., editors, *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 3982–3992, Hong Kong, China. Association for Computational Linguistics.
- Ruff, L., Vandermeulen, R., Goernitz, N., Deecke, L., Siddiqui, S. A., Binder, A., Müller, E., and Kloft, M. (2018). Deep one-class classification. In Dy, J. and Krause, A., editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 4393–4402. PMLR.
- Ruff, L., Vandermeulen, R. A., Görnitz, N., Binder, A., Müller, E., Müller, K.-R., and Kloft, M. (2020). Deep semi-supervised anomaly detection. In *International Conference on Learning Representations*.
- Sousa, R. F. d., Brum, H. B., and Nunes, M. d. G. V. (2019). A bunch of helpfulness and sentiment corpora in brazilian portuguese. In *Symposium in Information and Human Language Technology - STIL*. SBC.
- Souza, F., Nogueira, R., and Lotufo, R. (2020). Bertimbau: Pretrained bert models for brazilian portuguese. In Cerri, R. and Prati, R. C., editors, *Intelligent Systems*, pages 403–417, Cham. Springer International Publishing.
- Xu, H., Pang, G., Wang, Y., and Wang, Y. (2023a). Deep isolation forest for anomaly detection. *IEEE Transactions on Knowledge and Data Engineering*, 35(12):12591–12604.

Xu, Y., Gabor, K., Milleret, J., and Segond, F. (2023b). Comparative analysis of anomaly detection algorithms in text data. pages 1234–1245.

Identificação de aspectos explícitos e implícitos em críticas gastronômicas em português: avaliando o potencial dos LLMs

Luiz H. N. Silva¹, Eloize R. M. Seno¹, Rozane R. Rebechi², Helena M. Caseli³
Fabiano M. Rocha Júnior¹, Guilherme A. Faller²

¹Instituto Federal de São Paulo (IFSP) – São Carlos, SP

² Depto de Letras – Universidade Federal do Rio Grande do SUL – Porto Alegre, RS

³ Depto de Ciência da Computação – Universidade Federal de São Carlos – São Carlos, SP

{nascimento.henrique1, fabiano.j}@aluno.ifsp.edu.br, guilfaller@gmail.com
eloize@ifsp.edu.br, rozanereb@gmail.com, helenacaseli@ufscar.br

Abstract. *Aspect identification is a fundamental step in Aspect-Based Sentiment Analysis (ABSA), which involves detecting the opinion target aspects in product or service reviews published on social media. Although there are many works developed for detecting aspects in English, there are few studies in this area for Portuguese, and LLMs have been little explored. Given this context, this research investigated the potential use of LLMs for aspect identification in culinary reviews in Portuguese.*

Resumo. *A identificação de aspectos é uma etapa fundamental da Análise de Sentimentos Baseada em Aspectos (ASBA) que consiste em detectar os aspectos alvos de opinião em avaliações de produtos ou serviços publicadas nas mídias sociais. Enquanto existem vários estudos focados na detecção de aspectos na língua inglesa, para o português há poucos trabalhos na área e os LLMs praticamente não têm sido explorados. Dado esse contexto, esta pesquisa investigou o potencial de uso de LLMs na identificação de aspectos em críticas gastronômicas em português.*

1. Introdução

A análise de sentimentos baseada em aspectos (ASBA) é uma subárea da Análise de Sentimentos (AS) que busca identificar e analisar opiniões e sentimentos relacionados a aspectos ou atributos específicos de uma entidade, produto ou serviço. Em uma avaliação de um restaurante, por exemplo, aspectos como “comida”, “serviço” e “preço” podem ser analisados individualmente, permitindo uma compreensão mais detalhada das opiniões dos clientes sobre cada um deles.

A ASBA representa o nível mais complexo da análise automática, devido à dificuldade de se modelar as conexões semânticas entre um determinado aspecto (termo) e as palavras que fazem parte do seu contexto [Zhang et al. 2018]. Uma etapa fundamental da ASBA consiste na identificação de aspectos, os quais podem ser explícitos ou implícitos, de acordo com a literatura [Schouten and Frasinca 2016, Soni and Rambola 2022]. Enquanto o aspecto explícito ocorre diretamente no texto, o aspecto implícito não é mencionado explicitamente, mas pode ser inferido pelo contexto. Por exemplo, na avaliação de um restaurante “A comida estava deliciosa, mas demorou muito para chegar.”, temos um aspecto explícito “comida” com sentimento positivo e um aspecto implícito “serviço” com sentimento negativo.

Enquanto para o inglês há uma vasta literatura relacionada à detecção de aspectos [Schouten and Frasinicar 2016, Zhang et al. 2018, Soni and Rambola 2022], para o português as pesquisas ainda são emergentes [Pereira 2021]. Além disso, os trabalhos existentes se baseiam principalmente no uso de regras, léxicos e em algoritmos de aprendizado de máquina, sendo que o uso de modelos de linguagem em larga escala (*Large Language Model – LLM*, no inglês) tem sido pouco explorado. À medida que o interesse público por modelos generativos pré-treinados, como os modelos da OpenAI, continua a crescer, espera-se que a utilidade desses modelos em resolver tarefas de PLN seja investigada. E nesse sentido, algumas iniciativas recentes têm surgido [Oliveira et al. 2023, Santos and Paraboni 2023].

Dado esse contexto, este estudo investigou a potencialidade de cinco LLMs na identificação de aspectos explícitos e implícitos em críticas gastronômicas em português. Críticas gastronômicas são textos escritos por críticos profissionais da gastronomia com experiência em avaliar restaurantes, pratos e experiências culinárias. A escolha desse domínio se justifica pelo fato de que as críticas gastronômicas, até onde se sabe, ainda não foram exploradas no contexto da ASBA em português.

2. Trabalhos Relacionados

Os trabalhos de identificação de aspectos para o português se baseiam, principalmente, no uso de léxicos [Costa and Pardo 2020], regras de linguagem [Vargas and Pardo 2020, Machado et al. 2021], em algoritmos de aprendizado de máquina tradicionais [Balage Filho 2017, Vargas and Pardo 2018] e no uso de *deep learning* [Lopes et al. 2021, Assi et al. 2022, Machado and Pardo 2022, Resplande et al. 2022]). Em [Resplande et al. 2022], por exemplo, os autores avaliaram o uso de modelos baseados em *Transformers* na extração de aspectos em avaliações de hotéis. Os aspectos extraídos foram classificados, posteriormente, como positivos, negativos ou neutros usando o LLM GPT-3. Em um trabalho anterior [Seno et al. 2024], o GPT-3.5 Turbo foi empregado na tarefa de detecção de aspectos e classificação de polaridade em comentários do domínio político. Em [Machado 2023], os autores compararam o uso de LLMs – GPT-3.5, Maritaca e Llama – com um modelo BERT e com vários classificadores tradicionais na identificação de aspectos em revisões de produtos eletrônicos, livros e hotéis. Nos experimentos, os melhores resultados para os aspectos explícitos foram obtidos pelo classificador *CRF* (o melhor *F-score* foi 81% para revisões de hotéis). Porém, para os aspectos implícitos o melhor resultado, em termos de porcentagem de acerto, foi obtido com o Llama 7B (52%).

De forma similar, este estudo também explorou o uso do GPT-3.5 e dos modelos da família Maritaca na detecção de aspectos em críticas gastronômicas. Porém, os modelos investigados aqui são variações mais recentes das versões usadas por [Machado 2023].

3. Identificação de aspectos em Críticas Gastronômicas usando LLMs

Para a identificação de aspectos em críticas gastronômicas foram explorados alguns dos LLMs mais populares da atualidade como o GPT-3.5 Turbo, o GPT-4o e GPT-4o mini¹. Segundo a OpenAI², o GPT-4o é o seu modelo mais avançado e inteligente para tarefas

¹<https://platform.openai.com/docs/api-reference/introduction>

²<https://platform.openai.com/docs/models>

Tabela 1. Prompts usados na anotação de aspectos explícitos e implícitos.

<p>Aspectos explícitos: Dada a sentença EXEMPLO com os alvos de opiniões explícitos, identifique os alvos de opinião explícitos na sentença (se houver) no formato [e - alvo1], se não houver nenhum alvo, indique com um '-'. EXEMPLO: “A pizza estava gostosa. E a sobremesa também.”. Saída: [e - pizza] [e - sobremesa]</p>
<p>Aspectos implícitos: Dada a sentença EXEMPLO com os alvos de opiniões implícitos, identifique os alvos de opinião implícitos (se houver) no formato [i - alvo], se não houver nenhum alvo, indique com um '-'. EXEMPLO: “A pizza estava gostosa, mas era muito cara. Além disso, estava fria”. Saída: [i - preço] [i - temperatura]</p>

mais complexas. O GPT-4o mini é o modelo mais avançado na categoria de modelos pequenos, que também inclui o GPT-3.5 Turbo. Além desses LLMs, também foram investigados dois modelos monolíngues treinados para o português, o Sabiá-2-medium e o Sabiá-3³. Em experimentos reportados por [Almeida et al. 2024], o Sabiá-2-medium é comparado a vários outros LLMs, alcançando desempenho igual ou melhor que GPT-3.5 Turbo em várias análises. O Sabiá-3, por sua vez, lançado em julho de 2024, até o momento da escrita deste artigo não se tinha informações sobre o seu desempenho.

Todos os LLMs são modelos generativos baseados em *prompt*, que recebem como entrada um texto (*prompt*) contendo a descrição da tarefa a ser realizada e geram as saídas conforme solicitado. O grande desafio em lidar com esses modelos consiste em definir um *prompt* que gere as saídas exatamente como se espera para a tarefa. Vários *prompts* diferentes foram testados para a identificação de aspectos explícitos e implícitos no *corpus*. Foram experimentados *prompts* específicos para cada tipo de aspecto usando exemplos de anotação humana (i.e. abordagem *few-shot*) e sem o uso de exemplos de anotação (i.e. abordagem *zero-shot*). Contudo, percebeu-se uma facilidade maior dos modelos ao usar a abordagem *few-shot*. Assim, na anotação do *corpus* foram adotados os *prompts* apresentados na Tabela 1. Em todos os LLMs investigados a temperatura foi ajustada em zero, a fim de obter modelos mais determinísticos, conforme apontado por outros trabalhos da literatura [Oliveira et al. 2023, Santos and Paraboni 2023].

4. Corpus

Para os experimentos foi usado um conjunto de 1005 sentenças extraídas do *corpus* de críticas gastronômicas de [Rebecchi et al. 2021]. Cada sentença foi anotada por 5 anotadores humanos, todos pesquisadores da área de PLN, em duas etapas. Primeiramente os anotadores classificaram as sentenças em opinativa ou factual. Em seguida, aspectos explícitos e implícitos foram anotados, em dupla/trio, para as 374 (37,2%) sentenças consideradas opinativas pelos anotadores. Para estas, 432 aspectos foram identificados, sendo 88,6% explícitos e 11,4% implícitos. A Tabela 2 apresenta exemplos de sentenças com anotação de aspectos explícitos (em negrito) e implícitos.

Dado o fato de que não é possível determinar todos os aspectos possíveis para o *corpus*, não foi possível calcular o coeficiente *Kappa* para estimar a concordância entre os anotadores. Embora não se tenha obtido uma estimativa da concordância na anotação do *corpus*, a busca pelo consenso, seguida da clara convergência dos anotadores, permite assegurar que os aspectos identificados reproduzem de forma bastante fiel os aspectos que geralmente são considerados na avaliação de uma experiência gastronômica.

³Disponíveis por meio da MariTalk API como um *chatbot*.

Tabela 2. Exemplos de anotação de aspectos explícitos (em negrito) e implícitos.

Sentença	Implícito
Se estiver sozinho, desista de tentar o omakassê (sequência de iguarias decididas e enviadas aos poucos pelo chef) — ele é gigante (para uma pessoa) e caro (42 itens, R\$ 390).	tamanho; preço
Não é demais lembrar: a casa só aceita dinheiro ou cheque – costume fora de moda, também trazido de outros tempos.	forma de pagamento
Carta de vinhos: Excelente, com muitas opções argentinas para todos os bolsos.	variedade (vinhos); preço

5. Experimentos e Resultados

A Tabela 3 apresenta os resultados obtidos por cada LLM na detecção de aspectos explícitos e implícitos. O Sabiá-medium-2 obteve o melhor *F-score* (48,37%) para os aspectos explícitos, alcançando também a maior cobertura (77,75%). Contudo, a maior precisão (40,90%) foi obtida pelo GPT-4o mini. Já no que se refere aos aspectos implícitos, os resultados mostram uma grande dificuldade dos LLMs em identificar esse tipo de aspecto. Vale mencionar que essa dificuldade também foi relatada pelos humanos na anotação do *corpus*. Como os aspectos implícitos são inferidos pelo contexto, nem sempre é trivial perceber qual é o alvo de opinião. Em alguns casos, essa inferência exige um conhecimento mais especializado como no exemplo “Na boca, é equilibrado, com taninos firmes e boa estrutura.”, que se refere ao aspecto “vinho”. Para esse caso específico, apenas o modelo GPT-3.5 Turbo conseguiu identificar o aspecto implícito.

Tabela 3. Resultados obtidos para aspectos explícitos e implícitos.

LLM	Explícitos			Implícitos		
	Precisão	Cobertura	F-score	Precisão	Cobertura	F-score
Sabiá-2-medium	35,11%	77,75%	48,37%	1,60%	26,00%	3,01%
GPT-3.5 turbo	31,53%	76,18%	44,60%	1,95%	32,00%	3,68%
Sabiá-3	33,21%	67,80%	44,58%	2,21%	32,00%	4,13%
GPT-4o	21,51%	76,18%	33,54%	3,90%	34,00%	7,00%
GPT-4o mini	40,90%	15,82%	22,81%	2,23%	20,00%	4,01%

6. Conclusões

Este estudo investigou o uso de LLMs na detecção de aspectos em críticas gastronômicas. Nos experimentos, o LLM monolíngue Sabiá-2-medium mostrou um potencial maior na detecção de aspectos explícitos, do que os modelos multilíngues analisados. Enquanto que o Sabiá-3, também monolíngue, mostrou-se equivalente ao GPT-3.5 Turbo, superando o GPT-4o e o GPT-4o mini. Além de apresentarem desempenho superior ou equivalente aos obtidos pelos modelos multilíngues, os modelos monolíngues são bem mais acessíveis⁴. Com relação aos aspectos implícitos, todos os LLMs tiveram bastante dificuldade em identificar esse tipo de aspecto. O melhor desempenho foi obtido pelo GPT-4o (7% de *F-score*).

Como trabalhos futuros, pretende-se investigar a combinação de LLMs para a tarefa de identificação de aspectos, bem como a utilização de conhecimento do domínio de críticas gastronômicas para enriquecer os *prompts*.

⁴Os valores podem ser consultados em <https://openai.com/api/pricing/> e <https://www.maritaca.ai/>

Referências

- Almeida, T. S., Abonizio, H., Nogueira, R., and Pires, R. (2024). Sabiá-2: A new generation of portuguese large language models. *ArXiv*, abs/2403.09887.
- Assi, F. M., Candido, G. B., dos Santos Silva, L. N., Silva, D. F., and Caseli, H. M. (2022). Ufscar’s team at ABSAPT 2022: using syntax, semantics and context for solving the tasks. In Montes-y-Gómez, M. and et al., editors, *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2022)*, volume 3202 of *CEUR Workshop Proceedings*. CEUR-WS.org.
- Balage Filho, P. P. (2017). *Aspect extraction in sentiment analysis for portuguese language*. PhD thesis, São Carlos - SP.
- Costa, R. and Pardo, T. (2020). Métodos baseados em léxico para extração de aspectos de opiniões em português. In *Anais do IX Brazilian Workshop on Social Network Analysis and Mining*, pages 61–72, Porto Alegre, RS, Brasil. SBC.
- Lopes, E., Correa, U., and Freitas, L. (2021). Exploring BERT for aspect extraction in portuguese language. *The International FLAIRS Conference Proceedings*, 34.
- Machado, M., Pardo, T., Ruiz, E., and Felippo, A. (2021). Learning rules for automatic identification of implicit aspects in portuguese. In *Anais do XIII Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 82–91, Porto Alegre, RS, Brasil. SBC.
- Machado, M. and Pardo, T. A. S. (2022). Evaluating methods for extraction of aspect terms in opinion texts in Portuguese - the challenges of implicit aspects. In Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Odijk, J., and Piperidis, S., editors, *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3819–3828, Marseille, France. European Language Resources Association.
- Machado, M. T. (2023). *Methods for identifying aspects in opinion texts in Portuguese: the case of implicit aspects and their typological analysis*. PhD thesis, São Carlos - SP.
- Oliveira, A., Cecote, T., Silva, P., Gertrudes, J., Freitas, V., and Luz, E. (2023). How good is ChatGPT for detecting hate speech in portuguese? In *Anais do XIV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 94–103, Porto Alegre, RS, Brasil. SBC.
- Pereira, D. A. (2021). A survey of sentiment analysis in the portuguese language. *Artificial Intelligence Review*, 54(2):1087–1115.
- Rebechi, R. R., Nunes, R. R., Munhoz, L. R., and Marcon, N. O. (2021). Restaurant reviews in Brazil and the USA: A feast of cultural differences and their impact on translation. *Mutatis Mutandis. Revista Latinoamericana de Traducción*, 14:372–396.
- Resplande, J., Garcia, E., Junior, A., Rodrigues, R., Silva, D., Maia, D., Da Silva, N., Filho, A., and Soares, A. (2022). Deep learning Brasil at ABSAPT 2022: Portuguese transformer ensemble approaches. In Montes-y-Gómez, M. and et al., editors, *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2022)*, volume 3202 of *CEUR Workshop Proceedings*. CEUR-WS.org.

- Santos, W. and Paraboni, I. (2023). Predição de transtorno depressivo em redes sociais: Bert supervisionado ou ChatGPT zero-shot? In *Anais do XIV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 11–21, Porto Alegre, RS, Brasil. SBC.
- Schouten, K. and Frasincar, F. (2016). Survey on aspect-level sentiment analysis. *IEEE Transactions on Knowledge and Data Engineering*, 28(3):813–830.
- Seno, E., Silva, L., Anno, F., Rocha, F., and Caseli, H. (2024). Aspect-based sentiment analysis in comments on political debates in Portuguese: evaluating the potential of ChatGPT. In Gamallo, P., Claro, D., Teixeira, A., Real, L., Garcia, M., Oliveira, H. G., and Amaro, R., editors, *Computational Processing of the Portuguese Language: 16th Conference, PROPOR 2024*, pages 312–320, Santiago de Compostela, Galicia/Spain. Association for Computational Linguistics.
- Soni, P. K. and Rambola, R. (2022). A survey on implicit aspect detection for sentiment analysis: Terminology, issues, and scope. *IEEE Access*, 10:63932–63957.
- Vargas, F. A. and Pardo, T. A. S. (2018). Aspect clustering methods for sentiment analysis. In *Computational Processing of the Portuguese Language: 13th International Conference, PROPOR 2018, Canela, Brazil, September 24–26, 2018, Proceedings*, page 365–374, Berlin, Heidelberg. Springer-Verlag.
- Vargas, F. A. and Pardo, T. A. S. (2020). Linguistic rules for fine-grained opinion extraction. proceedings of the 14th International AAAI Conference on Web and Social Media, 2020.
- Zhang, L., Wang, S., and Liu, B. (2018). Deep learning for sentiment analysis : A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8.

Avaliação de Algoritmos de Clusterização para Agrupamento de Descrições de Produtos em Notas Fiscais Eletrônicas

Jonas Gabriel L. de Araújo¹, Thaís G. do Rêgo¹, Yuri de A. M. Barbosa¹

¹Centro de Informática – Universidade Federal da Paraíba (UFPB)
João Pessoa - PB - Brasil

jonas.araujo@academico.ufpb.br, gaudenciothais@gmail.com, yuri@ci.ufpb.br

Abstract. *The electronic invoice is essential for the tax audit process. This paper evaluates the effectiveness of clustering algorithms in grouping product descriptions from electronic invoices, a key document in tax audits. Due to the lack of standardization in these descriptions, clustering becomes a challenge. Using string similarity and adjustments for different units of measurement, DBSCAN, HDBSCAN, OPTICS, and Agglomerative Clustering were tested. Evaluation metrics included the Silhouette Coefficient, Calinski-Harabasz Index, and the percentage of grouped products. HDBSCAN showed the best initial performance, and the subclustering stage, while improving metrics, introduced inconsistencies in the groups.*

Resumo. *A nota fiscal eletrônica é essencial para o processo de auditoria fiscal. Este artigo avalia a eficácia de algoritmos de clusterização para agrupar descrições de produtos em notas fiscais eletrônicas, um desafio devido à falta de padronização nos registros. Usando similaridade de strings e ajustes para unidades de medida, foram testados DBSCAN, HDBSCAN, OPTICS e Agglomerative Clustering. As métricas de avaliação incluíram o Coeficiente de Silhueta, Índice de Calinski-Harabasz e a porcentagem de produtos agrupados. O HDBSCAN apresentou o melhor desempenho inicial, e a subclusterização, apesar de melhorar as métricas, introduziu inconsistências nos agrupamentos.*

1. Introdução

As Notas Fiscais Eletrônicas (NF-e) são um marco na modernização dos processos fiscais no Brasil, ao melhorar o controle e a fiscalização tributária, o que já resultou em avanços na arrecadação de impostos e no processo de auditoria [Vieira et al. 2019, Neto and Lopo Martinez 2016]. No entanto, a análise dessas notas enfrenta desafios devido à falta de padronização nas descrições de produtos, com erros ortográficos, abreviações e variações nas unidades de medida [Mazzarolo et al. 2022]. Essa inconsistência dificulta a organização e comparação de dados, exigindo técnicas computacionais para agrupar descrições similares e auxiliar na auditoria fiscal, que requer a correspondência entre o inventário das empresas e as notas emitidas por elas. Dessa forma, o uso de algoritmos de agrupamento facilita a fiscalização e melhora a eficiência do processo [Ribeiro et al. 2018].

Neste contexto, este estudo busca avaliar algoritmos de agrupamento, como DBSCAN [Ester et al. 1996], HDBSCAN [Campello et al. 2013], OPTICS [Ankerst et al. 1999] e *Agglomerative Clustering* (AGG) [Steinbach et al. 2000], para

agrupar descrições de produtos e identificar quais algoritmos oferecem o melhor desempenho na organização e interpretação dos dados. Para isso, foi empregada uma métrica personalizada no cálculo da matriz de distâncias, baseada em similaridade entre *strings* e a análise de uma segunda etapa de agrupamento dos dados.

2. Trabalhos relacionados

Nesta seção, serão abordados alguns trabalhos que contribuíram para tentar resolver as diferenças na padronização nas descrições dos produtos, a fim de melhorar o processo de fiscalização tributária no Brasil [Mazzarolo et al. 2022].

O trabalho de [Schulte et al. 2022] apresentou o ELINAC, um modelo que combina *autoencoder* e busca binária para agrupar descrições de produtos em notas fiscais. O método filtra as descrições, considerando apenas o nome e informações numéricas, como quantidade e dosagem. Embora eficiente, ele tem limitações ao distinguir produtos com variações sutis, como sabor.

A revisão de [Ahmed et al. 2022] aponta que a representação vetorial de textos curtos é desafiadora devido à alta dimensionalidade e ao ruído. O estudo de [Marinho et al. 2024] comparou representações textuais para classificar inconsistências em notas fiscais, calculando a similaridade entre a descrição do produto e a oficial da Nomenclatura Comum do Mercosul (NCM). Concluiu-se que a distância de edição de *strings* teve melhor desempenho preditivo do que *embeddings*, apesar de não considerar a similaridade entre produtos.

Este estudo se diferencia ao focar na avaliação de algoritmos de agrupamento e na representação de descrições de NF-es utilizando similaridade de *strings*. Enquanto outros trabalhos abordam redes neurais, detecção de fraudes e visualização de dados, este estudo explora a eficácia dos algoritmos de clusterização para organizar e interpretar as descrições de produtos em notas fiscais.

3. Metodologia

Esta seção descreve a base de dados, o cálculo da matriz de distâncias e os algoritmos de clusterização utilizados.

3.1. Base de dados

Foram usadas duas bases: uma base sintética com 22 descrições, contendo ruídos típicos [Mazzarolo et al. 2022], e uma base real cedida pela Secretaria da Fazenda da Paraíba (SEFAZ-PB) com 507 descrições. As descrições foram normalizadas, removendo caracteres especiais e convertendo tudo para caracteres maiúsculos.

3.2. Matriz de distâncias

Uma matriz de distâncias é uma matriz quadrada que contém as distâncias entre todos os pares de elementos do banco de dados. Neste trabalho, a matriz foi feita a partir de uma métrica personalizada, baseada na similaridade de Jaro [Jaro 1989]. O valor da similaridade varia entre 0 e 1, onde 0 indica que as *strings* não têm correspondências e 1 indica que as *strings* são idênticas. Entretanto, para o conceito de distância, quanto mais próximo de 0, mais próximos são dois pontos. Dessa forma, para computar a matriz de distância, foi calculado o complemento da similaridade de Jaro, ou seja, $1 - \text{JaroSimilarity}$.

Além da similaridade textual, foi introduzido um cálculo adicional para diferenciar produtos com o mesmo nome, mas com medidas distintas, como “200 ML” e “10 KG”. Isso evita que produtos com variação apenas na quantidade sejam considerados iguais. Para implementar esse ajuste, as medidas foram extraídas por meio da expressão regular 1 [Lucena et al. 2022], e convertidas em mililitros, gramas ou metros. Quando as medidas diferem, adiciona-se uma penalidade de 0,3 ao complemento da similaridade de Jaro, valor que foi escolhido após testes com variações entre 0,1 e 0,5.

$$(?:\d*[,]?\d+?\s?(?:kg|ml|mm|l|lt|gr|grs|g|metros|m|gb|k|cm|mg)\b) \quad (1)$$

3.3. Algoritmos de Clusterização

Para o agrupamento, este estudo avaliou 4 algoritmos diferentes: DBSCAN [Ester et al. 1996], HDBSCAN [Campello et al. 2013], OPTICS [Ankerst et al. 1999] e AGG [Steinbach et al. 2000]. Todos os algoritmos usados foram implementados pela biblioteca *scikit-learn*, versão 1.5.1, e nenhuma métrica de distância foi passada para os algoritmos, uma vez que a matriz já está pré-computada.

Os algoritmos foram usados em duas etapas: o agrupamento inicial e a *subclusterização* dos grupos de *outliers*, aplicada apenas na base real. Para o agrupamento inicial, foi definida uma distância máxima de agrupamento de 0,1 e um tamanho mínimo de *cluster* sendo igual a 2. Para a segunda etapa, a distância foi igual a 0,2. Os parâmetros de distância foram escolhidos após avaliação do agrupamento com variações entre 0,05 e 0,2 e os demais hiperparâmetros possuem os valores padrões da biblioteca.

Para avaliar o resultado dos agrupamentos, foram utilizadas duas métricas principais: o Coeficiente de Silhueta [Rousseeuw 1987], que avalia a coesão dos *clusters*, e o Índice de Calinski-Harabasz (CH) [Caliński and JA 1974], que mede a separação entre os grupos. O cálculo dessas métricas foi feito utilizando as distâncias entre pontos pré-computadas. Além disso, foi considerada a porcentagem de produtos agrupados para avaliar a cobertura dos dados pelos algoritmos de agrupamento.

4. Resultados e discussões

A Tabela 1 apresenta os resultados da primeira etapa dos experimentos. É importante ressaltar que o algoritmo AGG não gera um grupo de *outliers* identificado como -1 , o que exigiu um ajuste no cálculo das métricas para esse caso. Especificamente, todos os grupos individuais, que contêm apenas um produto, foram considerados como pertencentes ao grupo -1 , permitindo que as métricas fossem calculadas de forma consistente.

Na base sintética, DBSCAN, OPTICS e AGG produziram *clusters* idênticos, enquanto o HDBSCAN teve desempenho superior, distinguindo produtos com variações de sabor, mas não separando bem produtos de medidas diferentes. Nos dados reais, o HDBSCAN obteve as melhores métricas gerais, enquanto o OPTICS teve o maior coeficiente de Silhueta, mas o menor índice de CH, sugerindo que seus *clusters* não estavam bem separados.

A Tabela 2 apresenta os resultados da *subclusterização* dos grupos de produtos considerados *outliers* na base de dados real. Todas as métricas possuíram aumentos nos valores, quando comparados à primeira clusterização, especialmente na utilização do HDBSCAN, tanto na primeira, quanto na segunda etapa.

Tabela 1. Avaliação dos algoritmos de clusterização no agrupamento inicial

Base de Dados	Algoritmo	Silhueta	CH	Produtos agrupados (%)
Base Controlada	DBSCAN	0,490	7,97	98,16
	HDBSCAN	0,563	15,58	99,80
	OPTICS	0,490	7,97	98,16
	AGG	0,490	7,97	98,16
Base SEFAZ-PB	DBSCAN	0,686	21,71	86,19
	HDBSCAN	0,726	43,40	94,08
	OPTICS	0,730	17,98	85,99
	AGG	0,696	20,18	75,79

Embora as métricas tenham melhorado com a segunda etapa de clusterização usando o HDBSCAN, surgiram inconsistências nos agrupamentos. Por exemplo, produtos como “BOM TRIGO PREP. EMULSIF.” e “MARG. MEDALHA DE OURO” foram agrupados erroneamente no mesmo *cluster*. Isso indica que a fase adicional pode priorizar a melhoria das métricas, mas comprometer a consistência semântica, tornando os *clusters* menos úteis ou interpretáveis na prática.

Tabela 2. Avaliação dos algoritmos de clusterização no segundo agrupamento

Primeira Etapa	Segunda Etapa	Silhueta	CH	Produtos agrupados (%)
DBSCAN	DBSCAN	0,718	29,02	90,13
	HDBSCAN	0,737	79,30	97,63
	OPTICS	0,717	27,50	89,74
	AGG	0,719	28,06	89,94
HDBSCAN	DBSCAN	0,729	46,28	94,47
	HDBSCAN	0,740	125,55	99,21
	OPTICS	0,729	46,28	94,47
	AGG	0,729	46,28	94,47
OPTICS	DBSCAN	0,761	25,25	89,94
	HDBSCAN	0,779	73,11	97,63
	OPTICS	0,760	22,96	89,54
	AGG	0,763	23,46	89,74
AGG	DBSCAN	0,728	27,42	89,94
	HDBSCAN	0,746	75,52	97,43
	OPTICS	0,727	26,00	89,54
	AGG	0,729	26,53	89,74

5. Considerações finais

Este estudo avaliou os algoritmos de clusterização DBSCAN, HDBSCAN, OPTICS e AGG para agrupar descrições de produtos em NF-e, utilizando similaridade de *strings* como representação de dados. O HDBSCAN apresentou o melhor desempenho inicial, mas a segunda etapa de agrupamento gerou inconsistências. DBSCAN e OPTICS tiveram métricas um pouco inferiores, porém com menos irregularidades. Sugere-se, como trabalhos futuros, testar o método em bases maiores e explorar representações como *embeddings* e redes neurais para padronização.

Referências

- Ahmed, M., Tiun, S., Omar, N., and Sani, N. S. (2022). Short text clustering algorithms, application and challenges: A survey. *Applied Sciences*.
- Ankerst, M., Breunig, M. M., Kriegel, H.-P., and Sander, J. (1999). Optics: ordering points to identify the clustering structure. *SIGMOD Rec.*, 28(2):49–60.
- Caliński, T. and JA, H. (1974). A dendrite method for cluster analysis. *Communications in Statistics - Theory and Methods*, 3:1–27.
- Campello, R. J. G. B., Moulavi, D., and Sander, J. (2013). Density-based clustering based on hierarchical density estimates. In Pei, J., Tseng, V. S., Cao, L., Motoda, H., and Xu, G., editors, *Advances in Knowledge Discovery and Data Mining*, pages 160–172, Berlin, Heidelberg. Springer Berlin Heidelberg.
- Ester, M., Kriegel, H.-P., Sander, J., Xu, X., et al. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, volume 96, pages 226–231.
- Jaro, M. A. (1989). Advances in record-linkage methodology as applied to matching the 1985 census of tampa, florida. *Journal of the American Statistical Association*, 84(406):414–420.
- Lucena, L. F., de Menezes e Silva Filho, T., do Rêgo, T. G., and Malheiros, Y. (2022). Automatic recognition of units of measurement in product descriptions from tax invoices using neural networks. In Pinheiro, V., Gamallo, P., Amaro, R., Scarton, C., Batista, F., Silva, D., Magro, C., and Pinto, H., editors, *Computational Processing of the Portuguese Language*, pages 156–165, Cham. Springer International Publishing.
- Marinho, M., Weigang, L., Oliveira, V., and Borges, V. (2024). Estratégias computacionais baseadas em similaridade de textos e visualização exploratória para a identificação de inconsistências em notas fiscais eletrônicas.
- Mazzarolo, J., Steinmetz, R., and Mergen, S. (2022). Um estudo sobre a falta de padronização na descrição de produtos em notas fiscais eletrônicas. In *Anais da XVII Escola Regional de Banco de Dados*, pages 31–40, Porto Alegre, RS, Brasil. SBC.
- Neto, H. and Lopo Martinez, A. (2016). Nota fiscal de serviços eletrônica: Uma análise dos impactos na arrecadação em municípios brasileiros. *Revista de Contabilidade e Organizações*, 10:49.
- Ribeiro, L., Brandão, W., Marques, I., Andrade, P., Júnior, R., Oliveira, F., and Kelles, R. (2018). Reconhecimento de entidades nomeadas em itens de produto da nota fiscal eletrônica. 36:116–126.
- Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65.
- Schulte, J. P., Giuntini, F. T., Nobre, R. A., Nascimento, K. C. d., Meneguette, R. I., Li, W., Gonçalves, V. P., and Rocha Filho, G. P. (2022). Elinac: Autoencoder approach for electronic invoices data clustering. *Applied Sciences*, 12(6).
- Steinbach, M., Karypis, G., and Kumar, V. (2000). A comparison of document clustering techniques.

Vieira, P. A., Pimenta, D. P., Cruz, A. F. d., and Souza, E. M. S. d. (2019). Efeitos do programa de nota fiscal eletrônica sobre o aumento da arrecadação do estado. *Revista de Administração Pública*, 53(2):481–491.

Unified Knowledge-Graph for Brazilian Indigenous Languages: An Educational Applications Perspective

Gustavo Polleti¹, Fabio Cozman¹, Fabricio Gerardi²

¹Universidade de São Paulo, Brazil

²Universität Tübingen, Germany

gustavo.polleti@gmail.com

Abstract. *In this paper we present an unified knowledge-graph for Brazilian indigenous languages (BIL) from the perspective of potential applications, with a particular focus to the educational domain. We present BILGraph, a prototype we built for Bororo and Tupian languages, such as Guajajara, Munduruku and Akuntsu. Then we describe the knowledge extraction and entity linking process to build the graph from a dependency treebank and a lexical database for Tupian and Bororo languages. We discuss the limitations of BILGraph, highlighting ethical and practical implementation concerns.*

Resumo. *Este artigo apresenta um grafo de conhecimento unificado para as línguas indígenas brasileiras (BIL) a partir da perspectiva de aplicações potenciais, com foco particular no domínio educacional. Apresentamos o BIL-Graph, um protótipo construído para o Bororo e línguas tupis, como Guajajara, Munduruku e Akuntsu. Em seguida, descrevemos o processo de extração de conhecimento e ligação de entidades para construir o grafo a partir de um banco de árvores de dependências e de um banco de dados lexical para línguas Tupi e Bororo. Discutimos as limitações do BILGraph, destacando questões éticas e práticas de implementação.*

1. Introduction

The development of applications for Brazilian Indigenous languages (BIL) is severely limited by the lack of resources and tools. As is often the case with endangered languages, available resources are both scarce and dispersed [Pinhanez et al. 2023]. For some languages, such as Guajajara, Asurini, and Bororo, dictionaries are now available [Harrison and Harrison 2013, Cabral and Rodrigues 2003, Ferraz Gerardi]. For other languages, treebanks are available through the Universal Dependencies Project (UD) [Nivre et al. 2020a], though they vary in length and quality. Some languages, however, have only a handful of miscellaneous resources [Monserrat 2000]. This lack of standardization and proper linked data poses a significant barrier to developing tools and methods that could support language revitalization initiatives and accelerate the production of pedagogical material.

Recent efforts to unify Brazilian Indigenous language resources, such as TuLeD [Gerardi et al. 2022a] and the TuDeT treebanks on UD — a lexical database and a dependency treebank for several Tupian languages (still in their initial phase), respectively — have been pivotal in the development of language-learning applications targeted at Indigenous communities [Polleti 2024]. Additionally, the recent publication of

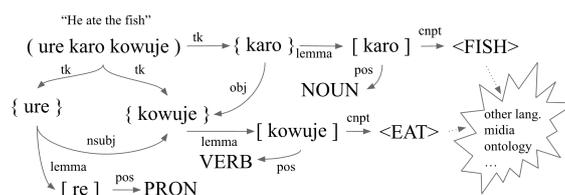


Figure 1. BILGraph toy example displaying a sampled subgraph associated with the sentence Bororo “ure karo kowuje”, i.e. “He ate the fish”.

the Bororo Corpus [Ferraz Gerardi et al. 2024], which is connected to the UD Treebank [Ferraz Gerardi 2024], has enabled the use of various computational tools to develop educational materials and other online resources; notably a language-learning app for the Bororo language.¹ UD-treebanks [Nivre et al. 2020b] are an important resource since the standardized type of annotation for all languages facilitate the development of new applications. On the other hand, heterogeneous and complex network structures, such as knowledge graphs, are known for their flexibility in incorporating linguistic characteristics [Cong and Liu 2014, Miller 1994] and can be effectively utilized to power sophisticated applications, including recommendation systems, information retrieval, and educational assistants.

In this work, we introduce a preliminary version of an unified knowledge graph for Brazilian indigenous languages, which we will refer as “BILGraph”, and we describe its knowledge extraction pipeline. We developed a prototype for Tupian languages available in Tuled and Tudet [Gerardi et al. 2022b], and the Bororo language [Ferraz Gerardi et al. 2024]. We discuss the knowledge graph prototype with a focus on the potential applications. We managed to develop a natural language processing pipeline to build BILGraph that can handle semi-structured data from several sources, such as annotated phrases from treebanks and dictionaries. We discuss the pipeline challenges and limitations. The main contribution of this work is to present a prototype version of BILGraph as a case of study on building an unified knowledge graph for BIL. We hope the knowledge graph and the methods presented in this work can support the development of sophisticated applications.

The paper is organized as follows. Section 2 describes BILGraph’s design and its development, including their data sources and knowledge extraction pipeline. Section 3 discusses the challenges and limitations of our prototype, analyses our processes and resources from both a practical implementation and potential applications perspective, and offers concluding remarks.

2. BilGraph: Linguistic Knowledge Graph

We have developed a knowledge extraction pipeline to structure and link language resources for Brazilian Indigenous Languages (BIL) available in Universal Dependencies (UD) treebanks and lexical databases, such as TuLeD and the Bororo dictionary. The result of this effort is “BILGraph”, a knowledge graph for BIL that contains four principal types of nodes: (1) sentence, (2) token, (3) lemma, and (4) concept. Consider the example depicted in Figure 1. The sentence node represents the Bororo treebank phrase *ure karo*

¹<https://bilingo-4388e.web.app/>

kowyje ‘He ate the fish’. This sentence node connects to its token nodes, which represent the individual words composing the sentence and their syntactic dependencies. In this example, “*kowyje*” is the root, with the object “*karo*” and the nominal subject “*ure*” linking to it. Each token node is connected to a single sentence node. Each token is further linked to a lemma node, which represents the word’s base form and its relationships to linguistic classes, including any applicable synonyms. Up to this point, the entities and relationships described are those typically found in dependency treebanks. However, the lexical database or dictionary adds another layer by linking lemma nodes to concept nodes. Concept nodes represent high-level abstractions that convey meaning across different languages and domains. In our example, the lemmas “*karo*” and “*kowyje*” are linked to the concepts “fish” and “eat,” respectively. The goal is to establish the concept nodes as a semantic layer that enables interoperability between the treebank sentences and other knowledge bases, such as ontologies, multimedia resources (e.g., phonetic or image databases), and other languages. Using BILGraph, one could easily search for sentences in other languages with similar structures or themes by fetching all sentence nodes connected to a given concept node. For example, a search engine could retrieve the Guajajara sentence *u?u ipiratetea?u* ‘It eats many fishes’ because it is connected to the concept node “FISH” as the similar sentence in Bororo *ure karo kowyje*. Note that the graph structure is flexible enough to encode N-N relationships between lemmas and concepts.

The relationships between sentences, tokens and lemmas can be extracted directly from UD treebanks, as the treebank sentences are annotated with attributes that allow a straightforward graph representation. To build BILGraph, the real challenge lies in linking lemma to concept nodes. In our preliminary version, we applied a simple entity linking process as follows. For each lemma, we generated a neighborhood set of similar words by changing and trimming characters based on rules. For example, in the Bororo language, we have different spellings where some words exchange “u” for “y”, and words like “boe” are often applied, so some of our neighborhood generation rules involved in adding or removing prefixes and changing exchangeable letters. The size of the neighborhood was defined considering a similarity threshold based on the Leveshnstein distance. Next, we select from all the vocabulary in our database the words that display high similarity, considering again a threshold based on leveshnstein distance, with at least one instance in our neighborhood. Finally, we test if dictionary entry or description for each candidate has at least one word in the sentence. So, for example, consider we are trying to link the lemma “*karo*”, from the sentence “He ate the fish”, to its appropriate concepts. Additionally, consider the dictionary description for a word candidate “*kabo*” is “a type of river fish”. In this case, we will establish the link due to the lexical similarity between “*karo*” and “*kabo*”, and due to the word “fish” that is present in both the dictionary entry and the sentence. Note that relying on lexical similarity may lead to inaccuracies. For example, the Bororo words “*apido*” (palm heart) and “*apodo*” (toucan) have high lexical similarity while their meanings are not related at all. If a dictionary entry contains both words, such as “palm hearth, edible for many animals like toucans”, this would lead to incorrect links being added to the graph. BILGraph’s knowledge extraction pipeline code, with the used Leveshentein distance thresholds for each language, and the knowledge graph itself is available in Github.² We adopted the RDF format, where each edge in the graph is represented as a triple.

²<https://github.com/gpadpoll/bilgraph>

3. Discussion and Concluding Remarks

The preliminary version of BILGraph introduced in this work represents a significant step forward in advancing resources for Brazilian Indigenous languages. We envision that BILGraph could power typical applications such as information retrieval from texts written in these languages, with a particular emphasis on its educational potential. The process of creating educational resources often involves organizing texts based on their linguistic characteristics, themes, and complexity levels. For instance, one might search for specific sentences to teach someone how to ask for food. BILGraph simplifies this task by allowing queries for sentences linked to specific concept nodes. To find sentences that include food-related vocabulary, one can attach a generic ontology to BILGraph’s concept nodes and search for sentences associated with food-related concepts. Moreover, BILGraph makes it easy to query sentences based on linguistic features, such as those using possessive pronouns, verb forms, plurals, adverbs, and more. We believe that BILGraph’s ability to query and organize sentences can enhance the use of treebanks and other available BIL resources in the development of educational materials. By organizing resources in a standardized and unified format, we can develop applications that scale across multiple languages. For example, a query that searches for food-related concepts in sentences for one language can be reused for other languages included in BILGraph. We are already leveraging BILGraph to develop a curriculum for a Bororo language course, which will be released as a language-learning app. We aim to extend this approach to other languages as they are incorporated into the knowledge graph.

At this point, our BILGraph prototype falls short in several aspects and remains a work in progress, from the difficulties of working with limited sources of data to inaccuracies and ethical concerns. BILGraph was built from TuLeD, TuDet and the Bororo treebank and dictionary. All these data sources were developed by compiling several sources from the literature, without a proper structured data gathering process. As a result, it suffers from incompleteness, notably when we consider coverage of dependency trees with translation to Portuguese. We only have Portuguese translations for “Bororo”, “Guajajara”, “Munduruku” and “Akuntsu” out of the 9 languages available. The lack of Portuguese translations limits the application of these resources, as for educational purposes for example. Furthermore, it is reasonable to expect that some inaccuracies may have been introduced as part of the entity linking and knowledge extraction process. We haven’t evaluated the correctness in a comprehensive manner yet, except for limited manual inspection by the researchers. Finally, it is worth mentioning ethical concerns. BILGraph has been developed without the involvement of indigenous community [Pinhanez et al. 2023], except for the case of Bororo, so it is hard to enforce ethical guidelines [Lewis et al. 2020], as for example proposed by the Los Pinos Declaration,³ before BILGraph can be properly inspected and validated by actual indigenous speakers.

We recognize a limitation in distinguishing similar forms that map to different lemmas. While various solutions exist, the most effective approach tend to be probabilistic, improving in accuracy with larger datasets. We also focus on further research in developing a pipeline which only uses the target language, without relying on the use of a dictionary. Overall, we hope BILGraph represents a positive step towards an unified source for BIL resources so that more tools and applications can be developed for them.

³<https://unesdoc.unesco.org/ark:/48223/pf0000374030>

Acknowledgements

The second author was partially supported by CNPq grant 305753/2022-3. We also thank support by CAPES -Finance Code 001. The authors of this work would like to thank the Center for Artificial Intelligence (C4AI-USP) and the support from the São Paulo Research Foundation (FAPESP grant 2019/07665-4) and from the IBM Corporation.

References

- Cabral, A. S. and Rodrigues, A. (2003). Dicionário da língua asurini do tocantins. *Belém-Pará: UFPA/IFNOPAP/UnB: IL/LALI*.
- Cong, J. and Liu, H. (2014). Approaching human language with complex networks. *Physics of Life Reviews*, 11(4):598–618.
- Ferraz Gerardi, F. *Bororo Dictionary*. Forthcoming. Available upon request.
- Ferraz Gerardi, F. (2024). *Universaldependencies/ud_bororo – bdt*.
- Ferraz Gerardi, F. M., Sollberger, D., and Toribio Serrano, L. (2024). *Corpus bororo (corbo) (v0.1.1)*.
- Gerardi, F. F., Reichert, S., Aragon, C., Wientzek, T., List, J.-M., and Forkel, R. (2022a). *TuLeD. Tupían Lexical Database*. Zenodo.
- Gerardi, F. F., Reichert, S., Aragon, C., Wientzek, T., List, J.-M., and Forkel, R. (2022b). *TuLeD. Tupían Lexical Database (v0.12)*.
- Harrison, C. and Harrison, C. (2013). *Dicionário Guajajara-Português*. SIL.
- Lewis, J. E., Abdilla, A., Arista, N., Baker, K., Benesiinaabandan, S., Brown, M., Cheung, M., Coleman, M., Cordes, A., Davison, J., Duncan, K., Garzon, S., Harrell, D. F., Jones, P.-L., Kealiikanakaoleohaililani, K., Kelleher, M., Kite, S., Lagon, O., Leigh, J., Levesque, M., Mahelona, K., Moses, C., Nahuewai, I. I., Noe, K., Olson, D., Parker Jones, Ō., Running Wolf, C., Running Wolf, M., Silva, M., Fragnito, S., and Whaanga, H. (2020). Indigenous protocol and artificial intelligence position paper. Project Report 10.11573/spectrum.library.concordia.ca.00986506, Aboriginal Territories in Cyberspace, Honolulu, HI. Edited by Jason Edward Lewis. English Language Version of "Ka?ina Hana ?Ōiwi a me ka Waihona ?Ike Hakuhi Pepa Kūlana" available at: <https://spectrum.library.concordia.ca/id/eprint/990094/>.
- Miller, G. A. (1994). WordNet: A lexical database for English. In *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*.
- Monserrat, R. F. (2000). *Vocabulário Amondawa-Português, Vocabulário e frases em Arara e Português, Vocabulário Gavião-Português, Vocabulário e frases em Karipuna e Português, Vocabulário e frases em Makurap e Português, Vocabulário e frases em Suruí e Português, Pequeno dicionário em Tupari e Português*. Universidade do Caixas do Sul.
- Nivre, J., Abrams, M., Agić, Z., Ahrenberg, L., Antonsen, L., Aranzabe, M. J., Arutie, A., Asahara, M., Ateyah, L., Attia, M., et al. (2020a). Universal dependencies v2: An evergrowing multilingual treebank collection. <https://universaldependencies.org/>. Accessed: 2024-08-27.

- Nivre, J., de Marneffe, M.-C., Ginter, F., Hajič, J., Manning, C. D., Pyysalo, S., Schuster, S., Tyers, F., and Zeman, D. (2020b). Universal Dependencies v2: An evergrowing multilingual treebank collection. In Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France. European Language Resources Association.
- Pinhanez, C. S., Cavalin, P., Vasconcelos, M., and Nogima, J. (2023). Balancing social impact, opportunities, and ethical constraints of using ai in the documentation and vitalization of indigenous languages. In Elkind, E., editor, *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23*, pages 6174–6182. International Joint Conferences on Artificial Intelligence Organization. AI for Good.
- Polleti, G. (2024). Building a language-learning game for Brazilian indigenous languages: A case study. Technical report, arXiv:2403.14515.

A Robustness Analysis of Automated Essay Scoring Methods

Rafael T. Anchiêta¹, Rogério F. de Sousa¹, Raimundo S. Moura²

¹Federal Institute of Piauí – (IFPI - Picos)
Laboratory of Artificial Intelligence, Robotics, and Automation (LIARA)

²Federal University of Piauí – (UFPI - Teresina)
Laboratory of Natural Language Processing (LPLN)

{rta, rogerio.sousa}@ifpi.edu.br, rsm@ufpi.edu.br

***Abstract.** This paper analyzed the robustness of a state-of-the-art Automated Essay Scoring (AES) model by applying various linguistically motivated perturbations to the Essay-BR corpus. Our findings reveal that the AES model failed to detect these adversarial modifications, often assigning higher scores to the disturbed essays than to the original ones.*

1. Introduction

Automated Essay Scoring (AES) aims to provide computational models for automatically grading essays or with minimal involvement of humans [Page 1966]. Although this research area is over fifty years old [Beigman Klebanov and Madnani 2020], it has recently gained the attention of the Brazilian community because of publicly available corpora [Marinho et al. 2021, Marinho et al. 2022a]. Several methods to grade an essay or its characteristics arose based on these resources [de Sousa et al. 2024, Oliveira et al. 2023, Marinho et al. 2022b]. Besides, there is a growing interest in the area. For instance, recently occurred the PROPOR’24 Competition, whose goal was to develop computer systems capable of automatically evaluating essays [Mello et al. 2024].

Despite the advances achieved, the Brazilian community has made little effort to evaluate the robustness of AES methods, including analyzing their sensitivity to adversarial perturbations. [Liu et al. 2024] define robustness as the capacity to remain stable and reliable under different circumstances. Studies demonstrate that AES methods for the English language are easily fooled [Perelman 2014], reducing the trustworthiness of AI-based automated scoring systems [Kabra et al. 2022]. Based on these limitations of AES methods for English, we investigated whether AES methods for Portuguese suffer from robustness problems.

Our objective is to analyze AES methods using adversarial essays. For that, we applied a set of perturbations to an essay corpus, including adding unrelated texts, shuffling, deleting, and repeating paragraphs of an essay. With these linguistically-motivated disturbances, we evaluated a state-of-the-art AES strategy for Portuguese and found that the analyzed model could not deal with adversarial essays, producing, in fact, better results for undisturbed essays.

The remainder of this paper is organized as follows: Section 2 briefly presents related work. In Section 3, we detailed the performed analysis to verify the robustness of an Automated Essay Scoring method for Portuguese. Finally, Section 4 concludes the paper and indicates future directions.

2. Related Work

[Kabra et al. 2022] proposed a model agnostic adversarial evaluation scheme and associated metrics for AES systems to test their natural language understanding capabilities and overall robustness. They evaluated models ranging from feature-engineering-based approaches to the latest deep-learning algorithms. The authors found that AES models are highly overstable such that even heavy modifications (as much as 25%) with content unrelated to the topic of the questions do not decrease the score produced by the models.

[Liu et al. 2024] evaluated Automatic Essay Scoring models’ robustness and generalization capabilities through a comprehensive series of experiments to validate various models’ efficacy. The authors randomly select a part of the essays and shuffle the order of the sentences or delete a sentence randomly to construct a Chinese adversarial sample set for evaluating the robustness of the models. The results showed that the advanced AES models have poor robustness and generalization ability, and Large Language Models have better performance but still need to be improved.

3. Robustness Analysis

The Essay-BR corpus [Marinho et al. 2021] is organized into training, development, and testing sets, each with 3,198, 686, and 686 essays. We used the test set to generate adversarial essays. First, we extracted the essays with a score greater than or equal to 680 since the average score of ENEM 2023 was 641.6 points¹, resulting in 305 essays. We adopted the strategy of selecting the best essays, avoiding those with several grammatical, structural, and argumentative issues. After that, we applied several perturbations to the essays to produce adversarial essays. From the original and adversarial essays set, we evaluated the robustness of an Automated Essay Scoring (AES) model.

We implemented linguistically motivated perturbations to analyze the robustness of an AES model, i.e., to check whether the model can detect any difference between original and modified responses. The perturbations are detailed below.

Add unrelated text. We added an unrelated paragraph in each essay. We create three sets of essays with unrelated content, each indicating the position where an unrelated text was added. The sets are with unrelated texts added at the beginning, middle, and end of the essays. We extracted a paragraph from essays with a prompt different from the analyzed essay and added it to the essay. This test tries to mimic the behavior of students when they make their responses lengthy by adding irrelevant information.

Add song and cake recipe. Although these perturbations add unrelated content to an essay, they have a very different language structure than written prose in essays. So, this can be used to test a system negatively. Furthermore, it has been observed that students use this strategy in their exams, possibly in an attempt to fool the system². We created two sets of perturbations, one for cake recipe and the other for the song. In both sets, we add unrelated content in the middle of the essays.

¹<https://querobolsa.com.br/revista/redacao-enem-2023-quantos-texto-s-tiraram-nota-mil-quantos-zeraram>

²<https://g1.globo.com/educacao/noticia/2013/03/queria-testar-correcao-do-enem-diz-jovem-que-pos-receita-na-redacao.html>

Add repeated text. For this adversarial strategy, we also created three sets of perturbations. For each set, we repeated the essay content at the beginning, middle, and end of the essays. The motivation for this perturbation is that, according to [Kabra et al. 2022], students sometimes repeat sentences or specific keywords in their responses to make them longer yet not out of context and to fashion cohesive paragraphs [Higgins and Heilman 2014, Yoon et al. 2018].

Delete text. Similar to adding repeated text, we created three sets of perturbations in this strategy. For each set, we removed a paragraph at the beginning, middle, and end of the essays. According to [Kabra et al. 2022], these tests generally break the flow of an argument, delete crucial details from an essay, and decrease wordiness. This perturbation can seriously detract from the coherency and quality of writing and frustrate readers.

Shuffle text. For this perturbation, we randomly shuffle the content of an essay. The motivation for this adversarial strategy is to analyze important aspects of essay scoring, such as coherence and organization, which measure the extent to which the response demonstrates a unified structure and direction of the narrative [Barzilay and Lapata 2008, Tay et al. 2018].

After generating adversarial essays, we evaluated a state-of-the-art automated essay scoring [de Sousa et al. 2024] based on the BERT model [Devlin et al. 2019]. We assessed that model using each ENEM competency through the Quadratic Weighted Kappa (QWK) metric [Cohen 1968] for original and adversarial essays. QWK is a metric commonly used to assess AES models [Yannakoudakis and Cummins 2015]. Table 1 shows the results on the original essays, and Table 2 presents the results on the adversarial essays.

Tables 1 and 2, from C1 to C5, indicate the five competencies of the ENEM, and the total is the final grade for an essay. In Table 2, we highlight the values greater than or equal to the value of the original essays.

Analyzing the values from the two tables, we can see that only the values of adding text at the beginning and adding a cake recipe were not greater than the original essay values, indicating that the AES model was able to identify perturbations in the essays, penalizing their scores. On the other hand, the scores for adding unrelated text in the middle, in the end, and a song were greater than or equal to the values for the original essays. An interesting finding is that, despite adding an unrelated text at the end of an essay, the C5 score was not penalized. Competency 5 of the ENEM is dedicated to elaborating a proposal to solve the problem. The proposal normally appears at the end of an essay, and the AES model could not detect the unrelated content added to an essay. More than that, the final grade of original and adversarial essays had the same QWK value, suggesting that the AES model failed to capture this perturbation.

For the perturbation of repeating a text in the essay, the AES model graded the original and adversarial essays with the same score, mainly in competence four. Competence 4 evaluates the superficial structure of the text, that is, how the sentences and paragraphs are linked through cohesive elements. This way, the AES model should negatively score such responses. Besides, the scores for repeating a text in the middle and at the end of an essay had the same value as the original essays.

Another interesting finding is that deleting some parts of the essay improves its grade in various competencies. As we can see, the scores for deleting a text in the essay

are greater than or equal to the scores of the original essays, including the final score. These results show that the AES model could not identify a break in the flow of an argument when essential parts of the essay were removed.

Finally, and perhaps the most interesting finding, is that shuffling the paragraphs of an essay produces better results than the original essays. This result demonstrates that the AES model could not determine the cohesion and coherence of the essays. That is, the AES model did not identify the transition between the lines of the essays, verifying disconnected ideas that change the meaning substantially.

The source code of the AES model and for generating adversarial essays are publicly available at <https://github.com/liara-ifpi/essay-robustness>.

Table 1. Quadratic Weighted Kappa results on the original essays.

C1	C2	C3	C4	C5	Total
0.44	0.23	0.29	0.24	0.62	0.46

Table 2. Quadratic Weighted Kappa results on the adversarial essays.

Adversarial strategy	C1	C2	C3	C4	C5	Total
Add unrelated text at the begging	0.41	0.14	0.15	0.17	0.57	0.40
Add unrelated text in the middle	0.42	0.18	0.24	0.20	0.62	0.44
Add unrelated text at the end	0.43	0.22	0.28	0.24	0.62	0.46
Add song	0.38	0.21	0.23	0.20	0.64	0.42
Add cake recipe	0.38	0.18	0.22	0.18	0.61	0.41
Repeat text at the begging	0.44	0.20	0.23	0.24	0.59	0.44
Repeat text in the middle	0.43	0.21	0.27	0.24	0.61	0.46
Repeat text at the end	0.43	0.23	0.28	0.24	0.62	0.46
Delete text at the begging	0.40	0.19	0.30	0.22	0.66	0.45
Delete text in the middle	0.40	0.23	0.29	0.25	0.66	0.46
Delete text at the end	0.41	0.23	0.29	0.24	0.63	0.46
Shuffle text	0.47	0.20	0.29	0.24	0.64	0.47

4. Conclusion

This paper presented a robustness analysis for automatic essay scoring focusing on the Portuguese language. We used the Essay-BR corpus, which is based on the ENEM competencies, to perform that analysis. Our strategy was to add several perturbations to produce adversarial essays, aiming to check if a state-of-the-art automated essay scoring model can detect any difference between original and modified responses. From the analysis, we have learned that the automated essay scoring model could not identify the perturbations in the essays, producing scores that were even greater than the original responses. We hope that this analysis sheds light on this research area and helps develop more robust strategies for automatically grading essays.

For future work, we intend to develop more perturbations and create a toolkit to facilitate the creation of adversarial essays.

References

- Barzilay, R. and Lapata, M. (2008). Modeling local coherence: An entity-based approach. *Computational Linguistics*, 34(1):1–34.
- Beigman Klebanov, B. and Madnani, N. (2020). Automated evaluation of writing – 50 years and counting. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7796–7810, Online. Association for Computational Linguistics.
- Cohen, J. (1968). Weighted kappa: nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin*, 70(4):213–220.
- de Sousa, R. F., Marinho, J. C., Neto, F. A. R., Anchiêta, R. T., and Moura, R. S. (2024). PiLN at PROPOR: A BERT-based strategy for grading narrative essays. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 2*, pages 10–13, Santiago de Compostela, Galicia/Spain. Association for Computational Linguistics.
- Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (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.
- Higgins, D. and Heilman, M. (2014). Managing what we can measure: Quantifying the susceptibility of automated scoring systems to gaming behavior. *Educational Measurement: Issues and Practice*, 33(3):36–46.
- Kabra, A., Bhatia, M., Singla, Y. K., Jessy Li, J., and Ratn Shah, R. (2022). Evaluation toolkit for robustness testing of automatic essay scoring systems. In *Proceedings of the 5th Joint International Conference on Data Science & Management of Data*, pages 90–99, Bangalore, India. Association for Computing Machinery.
- Liu, R., Wang, X., Liu, J., and Zhou, J. (2024). A comprehensive analysis of evaluating robustness and generalization ability of models in aes. In *Proceedings of the 7th International Symposium on Big Data and Applied Statistics*, pages 1–5, Beijing, China. IOP Publishing.
- Marinho, J. C., Anchiêta, R. T., and Moura, R. S. (2021). Essay-br: a brazilian corpus of essays. In *XXXIV Simpósio Brasileiro de Banco de Dados: Dataset Showcase Workshop, SBBD 2021*, pages 53–64, Online. SBC.
- Marinho, J. C., Anchiêta, R. T., and Moura, R. S. (2022a). Essay-br: a brazilian corpus to automatic essay scoring task. *Journal of Information and Data Management*, 13(1):65–76.
- Marinho, J. C., C., F., Anchiêta, R. T., and Moura, R. S. (2022b). Automated essay scoring: An approach based on enem competencies. In *Anais do XIX Encontro Nacional de Inteligência Artificial e Computacional*, pages 49–60, Campinas, Brazil. SBC.
- Mello, R. F., Oliveira, H., Wenceslau, M., Batista, H., Cordeiro, T., Bittencourt, I. I., and Isotanif, S. (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, Santiago de Compostela, Galicia/Spain. Association for Computational Linguistics.
- Oliveira, H., Ferreira Mello, R., Barreiros Rosa, B. A., Rakovic, M., Miranda, P., Cordeiro, T., Isotani, S., Bittencourt, I., and Gasevic, D. (2023). Towards explainable prediction of essay cohesion in portuguese and english. In *Proceedings of the 13th International Learning Analytics and Knowledge Conference*, pages 509–519, Arlington TX USA. Association for Computing Machinery.
- Page, E. B. (1966). The imminence of... grading essays by computer. *The Phi Delta Kappan*, 47(5):238–243.
- Perelman, L. (2014). When “the state of the art” is counting words. *Assessing Writing*, 21:104–111.
- Tay, Y., Phan, M., Tuan, L. A., and Hui, S. C. (2018). Skipflow: Incorporating neural coherence features for end-to-end automatic text scoring. In *Proceedings of the Thirty-second AAAI conference on artificial intelligence*, pages 5948–5955, New Orleans, Louisiana, USA. AAAI Press.
- Yannakoudakis, H. and Cummins, R. (2015). Evaluating the performance of automated text scoring systems. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 213–223, Denver, Colorado. Association for Computational Linguistics.
- Yoon, S.-Y., Cahill, A., Loukina, A., Zechner, K., Riordan, B., and Madnani, N. (2018). Atypical inputs in educational applications. In Bangalore, S., Chu-Carroll, J., and Li, Y., editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 3 (Industry Papers)*, pages 60–67, New Orleans - Louisiana. Association for Computational Linguistics.

Classificação de Notícias em Português Utilizando Modelos Baseados em Transferência de Aprendizagem e Transformers

Wagner Narde¹, João Mendanha², Henrique Barbosa⁴, Frederico Coelho⁴,
Bruno Santos³, Luiz Torres²

¹Grupo Energisa – Brasil

²Dept. de Computação e Sistemas – Universidade Federal de Ouro Preto (UFOP)

³Dept. de Ciência da Computação – Universidade Federal da Bahia (UFBA)

⁴Dept. de Engenharia Eletrônica – Universidade Federal de Minas Gerais (UFMG)

wagner.b.n@hotmail.com, bruno.ps@ufba.br, luiz.torres@ufop.edu.br

Abstract. *Fake news spreads faster on some social networks than regular news, which can have different ramifications, from influences on election outcomes to deaths due to incorrect treatments of diseases. This work aims to employ methods based on transfer learning and Transformer-based machine learning models to classify the veracity of tweets in the Portuguese language (Brazil pt-BR). To this aim, a reliably labeled database was created and opened to free access. The dataset relates posts extracted from X (formerly Twitter) and their proximity with facts or fake information. Five Transformer models were subsequently trained in Portuguese. The fine-tuned BERT model, initialized with pre-training on Portuguese text, achieved superior performance, yielding an accuracy of 95.1%.*

Resumo. *Fake news se espalha mais rápido em algumas redes sociais do que notícias regulares, o que pode ter diferentes consequências, desde influências nos resultados eleitorais até mortes devido a tratamentos incorretos de doenças. Este trabalho tem como objetivo empregar métodos baseados em aprendizado por transferência e modelos de aprendizado de máquina baseados em Transformers para classificar a veracidade de tweets na língua portuguesa (Brasil pt-BR). Para isso, foi criada uma base de dados confiável e rotulada, aberta para acesso gratuito. O conjunto de dados relaciona postagens extraídas do X (anteriormente conhecido como Twitter) e sua proximidade com fatos ou informações falsas. Subsequentemente, cinco modelos Transformer foram treinados em português. O modelo BERT ajustado, inicializado com pré-treinamento em textos em português, alcançou um desempenho superior, obtendo uma acurácia de 95.1%.*

1. Introdução

Os veículos de comunicação de grande circulação por muito tempo foram jornais, revistas, rádio e televisão. Hoje, notícias circulam através de vídeos no *YouTube*, portais de notícia e em redes sociais como *Facebook*, *X* (Antigo *Twitter*) e *WhatsApp*. Tornando a Internet, um dos principais meios de comunicação e consumo de notícias. No Brasil, 65% usam a Internet e suas aplicações como principais fontes de informação, nos Estados

Unidos 53% e, no mundo, o número estimado é de 62% [NegociosSC 2024, Gente 2024, DataReportal 2024, Data 2024]. O que, por um lado, demonstra que a Internet ampliou o acesso à informação, mas, por outro lado, também transformou a forma como as notícias são consumidas e compartilhadas.

A ascensão da Internet e, conseqüentemente, das redes sociais democratizaram a produção de notícias, permitindo que qualquer pessoa assuma o papel de produtor de conteúdo sem a supervisão tradicional de jornalistas. Este fenômeno pode ter impactado negativamente na qualidade das informações disseminadas, resultando em um aumento de notícias que propagam desinformação ou divulgam informações falsas [Reis et al. 2019, Vargas et al. 2021].

2. Proposta de Modelo

Neste trabalho, propomos um modelo baseado em transferência de aprendizagem, *transformers* e aprendizagem supervisionada para classificar textos em português nas redes sociais, com foco na plataforma X. Também criamos uma base de dados em português¹ (162 amostras e balanceada), que relaciona textos da plataforma X com sua veracidade, visando melhorar a detecção de *fake news* e promover a qualidade da informação nas redes sociais.

3. Metodologia

O primeiro passo foi a construção de uma base de dados contendo postagens de usuários da rede social X (Antigo *Twitter*). Foi utilizada a ferramenta *Get Old Tweets (GOT)* [Henrique 2018] para coletar *tweets* históricos, incluindo notícias falsas. Dessa forma, cada *tweet* foi analisado e classificado manualmente para determinar sua proximidade com o fato, assegurando que o conjunto de dados fosse rigoroso e preciso. Esse processo permitiu a criação de um conjunto de dados robusto e rotulado, essencial para o treinamento e validação eficazes do modelo de classificação de textos em português proposto. Em seguida, os textos coletados passam por uma fase de preparação, onde são inicialmente pré-processados e, posteriormente, rotulados. Após esse processo, os dados são ajustados para servirem como entradas adequadas para os modelos de aprendizagem de máquina.

3.1. Conjunto de dados: Coleta e Processamento

Este trabalho utiliza dados textuais em português extraídos da plataforma de rede social X. A plataforma permite a extração de informações através de sua *Application Programming Interface (API)*. Com a rede social selecionada, iniciou-se a coleta de dados para compor a base de dados. O processo de obtenção dos *tweets* consistiu em buscar no site de checagem de notícias verdadeiras ou falsas (LUPA²) e pesquisar por elas utilizando a ferramenta GOT [Henrique 2018]. Para isso, foram realizadas filtragens de notícias e alinhamento temporal aproximado para obtenção das postagens realizadas sobre a notícia verificada.

¹<https://github.com/WagnerNarde/ML-Transformers-Tweets-falsos>

²<https://lupa.uol.com.br/>

3.2. Seleção dos Modelos de Aprendizagem

Neste trabalho, foram adotados modelos de aprendizagem baseados em *Transformers*. Originalmente desenvolvido para tradução automática, o *Transformer* se destacou por sua capacidade de capturar relações de dependência de longo alcance de forma eficaz. Buscou-se por modelos *Transformers* que receberam pré-treinamento em português, visando aproveitar a transferência de aprendizagem. Como resultado, optou-se por ajustar os seguintes modelos: BERT base pré-treinado em português brasileiro por [Souza et al. 2020]; BERT base pré-treinado em 104 idiomas, incluindo português, por [Devlin et al. 2019]; RoBERTa pré-treinado por [Liu et al. 2019] com um *corpus* de 6,9 milhões de frases em português; XLM-R base, pré-treinado por [Conneau et al. 2020] em 100 idiomas, incluindo português; e, por fim, o modelo ELECTRA uncased [Clark et al. 2020], pré-treinado especificamente em português.

3.3. Treinamento

Para avaliar a capacidade de generalização do modelo, foi utilizado o método de validação cruzada com 10 partições (10-fold cross-validation). Este método divide o conjunto de dados em 10 sub-conjuntos. Cada subconjunto é usado uma vez como conjunto de teste, enquanto os restantes são usados como conjunto de treinamento. Esse processo é repetido 10 vezes, garantindo que cada amostra do conjunto de dados seja utilizada para testes ao menos uma vez. Esse procedimento não apenas melhora a capacidade de generalização do modelo, mas também fornece uma estimativa mais robusta do desempenho do modelo em dados não vistos.

4. Resultados

As métricas de avaliação incluem Acurácia, F1-Score, Precisão, Sensibilidade (Recall) e MCC.

Tabela 1. Resultados obtidos em cada modelo

Modelo	Épocas	Acurácia	F1	Precisão	Sensibilidade	MCC
ELECTRA (uncased)	10	0.864	0.848	0.883	0.824	0.720
RoBERTa pré-treinado em Português	7	0.901	0.897	0.852	0.962	0.812
XLM-R pré-treinado em multi-idiomas	9	0.903	0.898	0.883	0.922	0.804
BERT com Pré-treinamento em Português	6	0.944	0.955	0.944	0.971	0.887
BERT com Pré-treinamento Multi-idioma	10	0.914	0.918	0.900	0.944	0.825

Os resultados deste trabalho, apresentados na Tabela 1, mostram que cada modelo de aprendizado de máquina treinado para a classificação de notícias em português teve um desempenho variado, dependendo do número de épocas e das características do próprio modelo. O ELECTRA (*uncased*), treinado com 10 épocas, apresentou o desempenho mais baixo, o que pode ser atribuído à falta de diferenciação entre letras maiúsculas e minúsculas, bem como à qualidade dos pesos de pré-treinamento disponíveis. O modelo RoBERTa pré-treinado em Português, configurado com 7 épocas, superou o ELECTRA, beneficiando-se de uma arquitetura que captura melhor as nuances linguísticas do português e diferencia entre maiúsculas e minúsculas. O modelo XLM-R pré-treinado em multi-idiomas, com 9 épocas, demonstrou uma leve superioridade em relação ao RoBERTa em termos de acurácia e F1, aproveitando o conhecimento adquirido em múltiplos idiomas para melhorar a compreensão semântica. Já o BERT com Pré-treinamento

Multi-idioma, utilizando 10 épocas, mostrou robustez com acurácia e F1 acima de 0.9, destacando-se pela capacidade de transferir conhecimento linguístico de um corpus multilingue para o português. Por fim, o BERT pré-treinado em Português foi o modelo com melhor desempenho geral, utilizando apenas 6 épocas de treinamento. Este modelo se destacou na classificação correta das notícias, com acurácia, F1 e MCC superiores, evidenciando a eficácia do pré-treinamento específico em português e a importância do ajuste fino dos hiperparâmetros para maximizar a eficácia do modelo em tarefas específicas de classificação de texto.

5. Discussão

Os modelos ELECTRA *uncased* pré-treinado em Português e RoBERTa pré-treinado em Português apresentaram resultados abaixo do esperado, pode-se levantar a questão de que se tais modelos passaram pelo mesmo processo de pré-treinamento dos outros métodos. O modelo RoBERTa exige mais recursos computacionais comparado com o Bert, além de ser um aprimoramento do mesmo, portanto, melhores resultados eram esperados desse modelo. O modelo ELECTRA sendo um modelo *uncase*, esperava-se um desempenho abaixo dos outros classificadores pré-treinados exclusivamente em português. Ainda assim, acredita-se que o modelo não conseguiu generalizar bem o problema.

O XLM-R foi um modelo originalmente proposto para a tradução de idiomas, por isso ele está disponível em versão multi-idiomas pré-treinado em vários idiomas, inclusive português. Apesar do XLM-R não ter sido originalmente proposto para classificação de texto, ele obteve resultados melhores que o ELECTRA.

O Modelo BERT com Pré-treinamento em Português obteve acurácia e F1 superiores a todos os outros modelos, mostrando que o pré-treinamento em português feito por [Souza et al. 2020] foi muito eficiente e contribuiu positivamente para o bom desempenho do modelo. Os resultados preliminares mostraram que o modelo foi capaz de classificar notícias de uma base de dados relativamente pequena, bases de dados com poucas amostras é um desafio em algumas áreas, como na saúde.

6. Conclusões

Este trabalho apresentou uma abordagem para detecção de *tweets* falsos em português através de NLP. Além disso, foi criada e disponibilizada uma base de dados balanceada com *tweets* classificados de forma confiável. A base possibilitou o treinamento de modelos para detecção de notícias falsas. Sendo que o modelo BERT com 6 épocas foi o melhor comparado aos outros modelos testados.

7. Trabalhos Futuros

Na continuação do trabalho, pretendemos estender a avaliação comparativa com outros modelos estado da arte da literatura de classificação de texto baseados em aprendizado profundo. Pretende-se aumentar a base de dados com mais dados rotulados, mantendo a confiabilidade, e também buscar dados de outras fontes. Além de mostrar os resultados da classificação de notícias verdadeiras, planeja-se apresentar também os resultados de classificação das notícias falsas, assim como utilizar outras estratégias para o treinamento, como a validação cruzada com 5 partições.

Referências

- [Clark et al. 2020] Clark, K., Luong, M.-T., Le, Q. V., and Manning, C. D. (2020). ELECTRA: Pre-training text encoders as discriminators rather than generators. In *ICLR*.
- [Conneau et al. 2020] Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.
- [Data 2024] Data, P. (2024). Global social media users in 2024. Accessed: 2024-06-28.
- [DataReportal 2024] DataReportal (2024). Social media users 2024 (global data & statistics). Accessed: 2024-06-28.
- [Devlin et al. 2019] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (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.
- [Gente 2024] Gente, G. (2024). Pandemia e o consumo de notícias nas redes sociais. <https://gente.globo.com/pandemia-e-o-consumo-de-noticias-nas-redes-sociais/>. Acessado em 28 de junho de 2024.
- [Henrique 2018] Henrique, J. (2018). Get old tweets programatically. Repository on GitHub.
- [Liu et al. 2019] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.
- [NegociosSC 2024] NegociosSC (2024). O uso da internet, redes sociais e mídia no brasil em 2024. <https://www.negociossc.com.br/blog/o-uso-da-internet-redes-sociais-e-midia-no-brasil-em-2024/>. Acessado em 28 de junho de 2024.
- [Reis et al. 2019] Reis, J. C. S., Correia, A., Murai, F., Veloso, A., and Benevenuto, F. (2019). Supervised learning for fake news detection. *IEEE Intelligent Systems*, 34(2):76–81.
- [Souza et al. 2020] Souza, F., Nogueira, R., and Lotufo, R. (2020). BERTimbau: pretrained BERT models for Brazilian Portuguese. In *9th Brazilian Conference on Intelligent Systems, BRACIS, Rio Grande do Sul, Brazil, October 20-23 (to appear)*.
- [Vargas et al. 2021] Vargas, F., Benevenuto, F., and Pardo, T. (2021). Toward discourse-aware models for multilingual fake news detection. In *Proceedings of the Student Research Workshop Associated with RANLP 2021*, pages 210–218.

Beyond Single Models: Leveraging LLM Ensembles for Human Value Detection in Text

Diego Dimer Rodrigues, Mariana Recamonde-Mendoza, Viviane P. Moreira

¹Instituto de Informática – (UFRGS)

{ddrodrigues, mrmendoza, viviane}@inf.ufrgs.br

Abstract. *Every text may reflect its writer’s opinions, and these opinions, especially in political contexts, are often tied to specific human values that they either attain or constrain. Identifying these values can provide policymakers with deeper insights into the underlying factors that influence public discourse and decision-making. While current large language models (LLMs) have shown promise across various tasks, no single model may generalize sufficiently to excel in tasks like human value detection. In this work, we utilize data from the Human Value Detection task at CLEF 2024 and propose leveraging multiple ensembles of LLMs to enhance the identification of human values in text. Our results found that the ensemble models achieved higher F1 scores than all baseline models, suggesting that combining multiple models can offer performance comparable to very large models but at much lower memory requirements.*

1. Introduction

People can agree or disagree on numerous topics even when using the same information. These differences arise largely from their individual beliefs about what is worth striving for, a concept referred to as (human) *values*. Human values can conflict or align, leading to a wide range of opinions on controversial issues. This divergence is one of the reasons for the formation of different political parties, each representing the values of specific groups [Kiesel et al. 2022].

Given its significance, the study of human values spans multiple disciplines, including social sciences [Schwartz 1994] and formal argumentation [Bench-Capon 2003]. Researchers focused on various aspects, such as classifying values, detecting them in text, and understanding their societal impact. In computer science, there is a growing body of work dedicated to value detection and emotion recognition from text [Dellaert et al. 1996, Tariq et al. 2019, Ammanabrolu et al. 2022]. These tasks are challenging and yet have a broad spectrum of applications, such as aiding policymakers in gauging public sentiment, detecting political alignment, and more.

In this work, we aim to advance the field of human value detection by leveraging multiple ensembles of Large Language Models (LLMs) to identify these values in text and enhance model performance. We adopt the value taxonomy presented in [Schwartz et al. 2012], which categorizes values into two types for each value—attained and constrained. However, our task focuses solely on identifying the presence of a value in a sentence, so we sum the attained and constrained versions to determine whether a sentence contains a particular value. We conduct this study with a dataset from CLEF 2024. The data is highly imbalanced, making this a challenging classification problem.

2. Background and Related Work

Human Value Detection has recently gained attention, particularly as the focus of a shared task at CLEF 2024. This task aimed to detect human values in speech, attracting participation from 20 teams. The outcomes of this competition, including the performance metrics of each team, are detailed in [Kiesel et al. 2022]. These efforts underscore the complexity of detecting nuanced human values in text and highlight the need for advanced models that can accurately capture such subtleties.

LLMs have revolutionized NLP tasks across various domains. The introduction of Transformer architectures [Vaswani et al. 2017] marked a significant leap forward, leading to the development of powerful pre-trained models like BERT [Devlin et al. 2019], RoBERTa [Liu et al. 2019], and DeBERTa [He et al. 2021]. These models have been highly effective in text classification, sentiment analysis, and content generation, significantly reducing the need for training models from scratch. Numerous studies [Xian et al. 2023, Hoang et al. 2019, Sun et al. 2019, Sobhanam and Prakash 2023] have demonstrated the efficacy of fine-tuning these models for specific tasks, showcasing their versatility and robustness in handling diverse NLP challenges.

Ensemble Learning is a well-established technique in machine learning, often employed to improve predictive performance by combining multiple models. Traditionally associated with decision trees [Quinlan 1986], ensemble learning has evolved to incorporate various frameworks, including those involving LLMs [Jiang et al. 2023].

3. Methodology

The data used in this study comes from the Human Value Detection at CLEF (Conference and Labs of the Evaluation Forum) 2024 task (ValueEval'24) [Kiesel et al. 2024a] and consists of approximately 3K human-annotated texts containing over 73K sentences. The annotation associated with each sentence indicates whether a specific human value is “attained” and “constrained”. A total of 19 human values are analyzed. Each column receives the value 0, 0.5, or 1, indicating whether the sentence does not contain the human value, partially contains it, or fully contains it, respectively. This study focused on the English dataset. All models were optimized for F1-Macro score.

To approach the task as a multi-label classification problem, we combined the “attained” and “constrained” columns in the *labels* file, summing their values to determine whether a specific human value is present in a sentence (0 for false, 1 for true). The result was an array of 19 boolean values for each sentence, which were then used as inputs for model fine-tuning. Thus, each human value represents a class and the predictive model may assign more than one class for a given sentence. While the value *Humility* was removed by many CLEF participants due to its scarcity in the training set (present in only 0.2% of sentences), we retained it, considering it important to predict even rare values to ensure comprehensive performance across all values.

Using the training dataset, we fine-tuned six models: base and large versions of BERT [Devlin et al. 2019], RoBERTa [Liu et al. 2019], and DeBERTa [He et al. 2021]. After fine-tuning, we used the validation data to create a new dataset that included the sentences, prediction probabilities for each class, and binary predictions indicating whether a value is present in a sentence. The true labels are also carried onto the dataset to enable evaluations. Five different ensemble approaches were used to combine model outputs:

- **prob-equal**: Probabilities from each model were summed and then averaged. A threshold of 0.2 was applied.
- **prob-large-double**: Probabilities from base models were summed, and probabilities from large models were doubled before summing. The total was divided by the number of votes (nine), and a threshold of 0.2 was applied.
- **preds-majority**: Binary predictions from all models were summed, with a threshold of 2 applied to predict a value as present if at least two models identified it.
- **preds-large-double**: Binary predictions were summed, with large models receiving two votes each. A threshold of 2 was used, meaning a value would be predicted as present if one large model or two base models identified it.
- **prob-weight-macro-f1**: The probabilities predicted by each model were weighted by their F1 scores on the validation set. The weighted probabilities were then summed and normalized, followed by applying a threshold of 0.2.

For reproducibility, all experiments, ensemble diagrams, and scripts used for fine-tuning are available on GitHub¹, with a fixed random seed for all libraries. Implementation details and further results are also in our repository. The models used in this study are publicly accessible and can be downloaded from HuggingFace.

4. Results

Results are presented in Table 1. The RoBERTa Large model achieved the highest accuracy among the individual models, which aligns with expectations given the larger model size. However, since the primary metric for model selection during training was the macro F1-score rather than accuracy, it is not surprising that larger models and ensemble models do not consistently show higher accuracy.

Table 1. F1 and Accuracy results for our models and baselines. * means the model is an ensemble, and † means it used the multilingual dataset version

	Model	Macro F1	Accuracy
Base models	BERT-base-uncased	0.160	0.502
	BERT-large	0.263	0.482
	RoBERTa-base	0.248	0.485
	RoBERTa-large	0.282	0.508
	DeBERTa-base	0.274	0.480
	DeBERTa-large	0.295	0.507
Ensembles	prob-equal	0.330	0.447
	prob-large-double	0.326	0.438
	prob-weight-macro-f1	0.330	0.445
	preds-majority	0.318	0.484
	preds-large-double	0.319	0.418
Baselines	[Legkas et al. 2024] †	0.390	–
	[Yunis 2024] * †	0.350	–
	[Yeste et al. 2024]	0.280	–

Table 1 also compares our results with the top-3 models from the CLEF 2024 submissions. Notably, our ensemble approaches, specifically *prob-weight-macro-f1* and

¹<https://github.com/diegodimer/valueeval24>

prob-equal, performed only 0.03 and 0.02 below the top-scoring models from the conference, which utilized XLM models and the multilingual dataset. The approach by Arthur Schopenhauer [Yunis 2024] leveraged an ensemble of DeBERTa-v2-xxlarge and xlmRoBERTa-large models. Similarly, Hierocles of Alexandria [Legkas et al. 2024] employed both the multilingual and English-translated datasets, incorporating sentence sequence information and fine-tuning an XLM-RoBERTa-xl model. Finally, team Philo of Alexandria [Yeste et al. 2024] fine-tuned a DeBERTa model specifically for this task.

Looking into the scores for each of the 19 human values, we see that our ensembles demonstrated competitive performance, closely matching the results of XLM models and outperforming the DeBERTa-base model across nearly all values. This task was particularly challenging due to the significant class imbalance in the dataset, with nearly 50% of test set instances not containing any of the 19 values. This imbalance skews predictions towards false negatives, resulting in lower F1 scores despite high accuracy, as models may correctly predict the absence of values due to their prevalence.

Overall, the results demonstrate that ensemble models can achieve performance comparable to very large models, even when utilizing models that require less computational resources. Although training an XLM-DeBERTa model was not feasible on the hardware used for this study due to memory constraints, our ensembles still achieved a strong macro F1-score. Specifically, the best ensemble model improved the macro F1-score from 0.295 (the highest among the base models) to 0.33, highlighting the effectiveness of ensemble methods in enhancing model performance in this context.

5. Conclusion

In this study, we tackled the complex task of identifying human values in text, a challenge crucial for understanding the values that shape public discourse and decision-making. By leveraging multiple ensembles of LLMs, we demonstrated that ensemble-based approaches could significantly enhance individual model performance in this task. This suggests that instead of relying solely on a single, powerful LLM, ensemble methods offer a more robust and effective solution for complex NLP tasks.

Despite the advanced capabilities of models like GPT-4.0, these models still struggle to consistently deliver satisfactory performance in this domain. For instance, in the ValueEval'24, a team using GPT-4.0 for zero-shot classification achieved an F1-score of 0.25 [Kiesel et al. 2024b], which is lower than the performance of our ensemble approaches. This highlights the inherent challenges in human value detection, where the nuances of language and context often exceed the capacity of a single model, no matter how sophisticated. Future work will include a qualitative analysis to better understand the errors made by the models and improve the proposed approaches, reinforcing the potential of ensemble learning as a key strategy in advancing the field.

Acknowledgments. This work has been partially funded by CNPq-Brazil and Capes Finance Code 001.

References

- [Ammanabrolu et al. 2022] Ammanabrolu, P., Jiang, L., Sap, M., Hajishirzi, H., and Choi, Y. (2022). Aligning to social norms and values in interactive narratives. In Carpuat,

- M., de Marneffe, M.-C., and Meza Ruiz, I. V., editors, *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5994–6017, Seattle, United States. Association for Computational Linguistics.
- [Bench-Capon 2003] Bench-Capon, T. J. M. (2003). Persuasion in practical argument using value-based argumentation frameworks. *Journal of Logic and Computation*, 13(3):429–448.
- [Dellaert et al. 1996] Dellaert, F., Polzin, T., and Waibel, A. (1996). Recognizing emotion in speech. In *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP '96*, volume 3, pages 1970–1973 vol.3.
- [Devlin et al. 2019] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding.
- [He et al. 2021] He, P., Liu, X., Gao, J., and Chen, W. (2021). DEBERTA: Decoding-enhanced BERT with disentangled attention. In *International Conference on Learning Representations*.
- [Hoang et al. 2019] Hoang, M., Bihorac, O. A., and Rouces, J. (2019). Aspect-based sentiment analysis using BERT. In Hartmann, M. and Plank, B., editors, *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 187–196, Turku, Finland. Linköping University Electronic Press.
- [Jiang et al. 2023] Jiang, D., Ren, X., and Lin, B. Y. (2023). Llm-blender: Ensembling large language models with pairwise ranking and generative fusion.
- [Kiesel et al. 2022] Kiesel, J., Alshomary, M., Handke, N., Cai, X., Wachsmuth, H., and Stein, B. (2022). Identifying the Human Values behind Arguments. In Muresan, S., Nakov, P., and Villavicencio, A., editors, *60th Annual Meeting of the Association for Computational Linguistics (ACL 2022)*, pages 4459–4471. Association for Computational Linguistics.
- [Kiesel et al. 2024a] Kiesel, J., Çöltekin, Ç., Heinrich, M., Fröbe, M., Alshomary, M., De Longueville, B., Erjavec, T., Handke, N., Kopp, M., Ljubešić, N., Meden, K., Mirzakhmedova, N., Morkevičius, V., Reitis-Münstermann, T., Scharfbillig, M., Stefanovitch, N., Wachsmuth, H., Potthast, M., and Stein, B. (2024a). Overview of touché 2024: Argumentation systems. In Goharian, N., Tonello, N., He, Y., Lipani, A., McDonald, G., Macdonald, C., and Ounis, I., editors, *Advances in Information Retrieval*, pages 466–473, Cham. Springer Nature Switzerland.
- [Kiesel et al. 2024b] Kiesel, J., Çöltekin, Ç., Heinrich, M., Fröbe, M., Alshomary, M., Longueville, B. D., Erjavec, T., Handke, N., Kopp, M., Ljubešić, N., Meden, K., Mirzakhmedova, N., Morkevičius, V., Reitis-Münstermann, T., Scharfbillig, M., Stefanovitch, N., Wachsmuth, H., Potthast, M., and Stein, B. (2024b). Overview of Touché 2024: Argumentation Systems. In Goeriot, L., Mulhem, P., Quénot, G., Schwab, D., Nunzio, G. M. D., Soulier, L., Galuscakova, P., Herrera, A. G. S., Faggioli, G., and Ferro, N., editors, *Experimental IR Meets Multilinguality, Multimodality, and Interaction. 15th International Conference of the CLEF Association (CLEF 2024)*, Lecture Notes in Computer Science, Berlin Heidelberg New York. Springer.

- [Legkas et al. 2024] Legkas, S., Christodoulou, C., Zidianakis, M., Koutrintzes, D., Petasis, G., and Dagioglou, M. (2024). Hierocles of alexandria at touché: Multi-task & multi-head custom architecture with transformer-based models for human value detection. In *Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024)*, *CEUR Workshop Proceedings*, *CEUR-WS.org*.
- [Liu et al. 2019] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach.
- [Quinlan 1986] Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, 1(1):81–106.
- [Schwartz 1994] Schwartz, S. H. (1994). Are there universal aspects in the structure and contents of human values? *Journal of Social Issues*, 50(4):19–45.
- [Schwartz et al. 2012] Schwartz, S. H., Cieciuch, J., Vecchione, M., Davidov, E., Fischer, R., Beierlein, C., Ramos, A., Verkasalo, M., Lönnqvist, J.-E., Demirutku, K., Dirilen-Gumus, O., and Konty, M. (2012). Refining the theory of basic individual values. *Journal of Personality and Social Psychology*, 103(4):663–688.
- [Sobhanam and Prakash 2023] Sobhanam, H. and Prakash, J. (2023). Analysis of fine tuning the hyper parameters in RoBERTa model using genetic algorithm for text classification. *International Journal of Information Technology*, 15(7):3669–3677.
- [Sun et al. 2019] Sun, C., Qiu, X., Xu, Y., and Huang, X. (2019). How to fine-tune BERT for text classification? In Sun, M., Huang, X., Ji, H., Liu, Z., and Liu, Y., editors, *Chinese Computational Linguistics*, pages 194–206, Cham. Springer International Publishing.
- [Tariq et al. 2019] Tariq, Z., Shah, S. K., and Lee, Y. (2019). Speech emotion detection using iot based deep learning for health care. In *2019 IEEE International Conference on Big Data (Big Data)*, pages 4191–4196.
- [Vaswani et al. 2017] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. *CoRR*, abs/1706.03762.
- [Xian et al. 2023] Xian, G., Guo, Q., Zhao, Z., Luo, Y., and Mei, H. (2023). Short text classification model based on DeBERTa-DPCNN. In *2023 4th International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE)*, pages 56–59.
- [Yeste et al. 2024] Yeste, V., Ardanuy, M., and Rosso, P. (2024). Philo of alexandria at touché: A cascade model approach to human value detection. In *Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024)*, *CEUR Workshop Proceedings*, *CEUR-WS.org*.
- [Yunis 2024] Yunis, H. (2024). Arthur schopenhauer at touché 2024: Multi-lingual text classification using ensembles of large language models. In *Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024)*, *CEUR Workshop Proceedings*, *CEUR-WS.org*.

Performance in a dialectal profiling task of LLMs for varieties of Brazilian Portuguese

Raquel M. Ko Freitag^{1,2}, Túlio Sousa de Gois^{1,3}

¹Laboratório Multiusuário de Informática e Documentação Linguística
Universidade Federal de Sergipe – Brazil (UFS)
Didática II – 49.107-230 – São Cristóvão – SE – Brazil

²Departamento de Letras Vernáculas – Universidade Federal de Sergipe – UFS

³Departamento de Computação – Universidade Federal de Sergipe – UFS

{rkofreitag,tuliosg}@academico.ufs.br

Abstract. *Different of biases are reproduced in LLM-generated responses, including dialectal biases. A study based on prompt engineering was carried out to uncover how LLMs discriminate varieties of Brazilian Portuguese, specifically if sociolinguistic rules are taken into account in four LLMs: GPT 3.5, GPT-4o, Gemini, and Sabiá-2. The results offer sociolinguistic contributions for an equity fluent NLP technology.*

Resumo. *Vieses de diferentes tipos são reproduzidos em respostas geradas por LLMs, inclusive dialetais. Um estudo baseado em engenharia de prompt foi realizado para descobrir como os LLMs discriminam as variedades do português brasileiro, especificamente se regras sociolinguísticas são consideradas por quatro LLMs – GPT 3.5, GPT-4o, Gemini e Sabiá-2 – na geração de suas respostas. Os resultados oferecem contribuições sociolinguísticas para uma tecnologia de PLN com equidade dialetal.*

1. Introduction

Advances in generative AI have enabled near-human responses, crucial for overcoming the Turing test [Danziger 2018]. However, achieving this requires algorithms to replicate ethically questionable human behaviors, including biases learned by large language models (LLMs) [Freitag 2021].

Biases can be explicit, consciously manipulated, or implicit, operating unconsciously through automatic associations. These biases affect generative AI in two key areas: the rules and filters applied during LLM fine-tuning, and the linguistic datasets used for training. However, the specifics of these biases—whether in rules, filters, or dataset selection—remain unclear [Bender et al. 2021]. To investigate these biases, reverse-engineering through prompt engineering is necessary, similar to how sociolinguistics studies human language attitudes.

In Brazil, sociolinguistic studies over the past 50 years have highlighted significant asymmetries between prestigious and non-standard varieties (whether regional or socially stigmatized), often perpetuated by implicit biases in educational materials and media, such as the portrayal of regional accents in soap operas [Freitag 2016]. These societal biases likely extend to LLMs.

For AI to be ethically and socially sensitive, the diversity of societal communities must be reflected in the language samples used to train LLMs. [Grieve et al. 2024] define a language variety as “a population of texts defined by external factors, such as the social background of the people who produce these texts, the social context in which these texts are produced, and the time period over which these texts are produced.”

Currently, there is no transparency on how language samples are collected and balanced to reflect linguistic diversity. A study using prompt engineering could reveal how LLMs handle varieties of Brazilian Portuguese and whether they consider sociolinguistic rules.

2. Dialectal biases in LLMs

Brazil’s continental size contributes to its dialectal diversity, further enriched by social diversity in language use. Both geolinguistic and sociolinguistic approaches have systematically described these patterns in Brazilian Portuguese [Roncarati et al. 2003, Abraçado and Martins 2015].

Though linguistic diversity doesn’t align strictly with socio-political boundaries, it is socially perceived and manifests in stereotypes, such as the classic *biscoito* vs. *bolacha* ‘cookie’ debate,¹ regional jokes and memes,² or even humorous maps that reflect aspects of perceptual dialectology [Preston 2010, Freitag et al. 2015, Freitag et al. 2016].



Figure 1. *Biscoito* or *bolacha* [A] and mineiros’s memes [B]

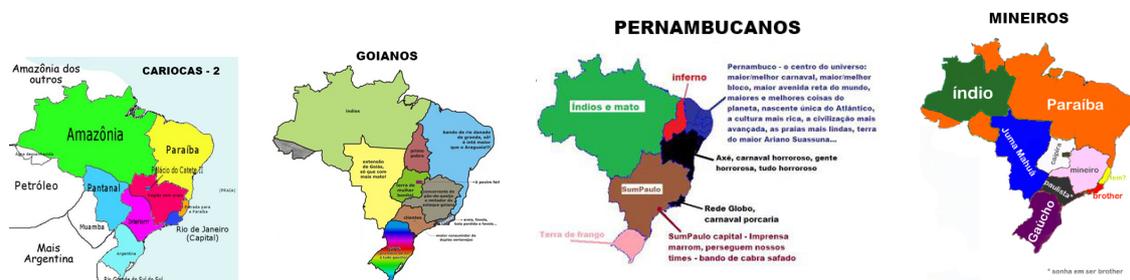


Figure 2. Funny maps with Brazilian stereotypes

Sociolinguistic surveys support these representations, showing systematic patterns between dialectal groups and territories, such as variations in first-person plural (*nós* a

¹<https://rionoticias.com.br/afinal-bolacha-ou-biscoito/>

²<https://www.conhecaminas.com/2016/10/10-coisas-que-so-mineiro-fala-entende-e.html>

gente) and second-person pronouns (*tu* *você*), their agreement (*nós vamos* *nós vai* — *tu vais* *tu vai*), and the shift between subjunctive and indicative moods for imperative forms (*cante* *canta*) [Abraçado and Martins 2015]. Phonological variations, like diminutives and diphthongs, also appear in written representations, particularly in memes.

It is expected that LLMs, like humans, will learn and reproduce societal biases and stereotypes, including linguistic biases [Shrawgi et al. 2024, Fleisig et al. 2024], as observed in LLMs trained on African American English [Mengesha et al. 2021, Dacon and Tang 2021, Dacon et al. 2022, Dacon 2022]. This paper evaluates the accuracy of LLMs in identifying Brazilian Portuguese dialectal profiles, assessing both their agreement with human judgments and consistency across different models: GPT-4o and GPT 3.5 (OpenAI), Sabiá-2 (Maritaca AI), and Gemini (Google AI). None of these LLMs disclose the size of their training corpus or the sources of their texts.

3. Method

The procedures encompassed three stages: generation, classification, and data wrangling. The dataset and analysis scripts are available at <https://osf.io/un8cw/>.

3.1. Target-profiles generation

The initial step involved instructing LLMs to generate text passages aimed at creating typical linguistic profiles for each of Brazil's 27 states. The prompt used was:

“Write a Facebook biography for my T-shirt store in XXX language, replacing XXX with the name of the state.”

Data was collected on 10 June 2024 and then analyzed for linguistic dialectal features to identify dialectal clues.

3.2. Target-profiles classification

LLMs were asked to identify the state of the text generated in the first step (*task*) by two prompts:

- *task + input*
*“This is a target audience identification task. Looking at the linguistic marks, identify for target audience in which state of Brazil the INPUT advertisement was constructed:
INPUT: XXXX”*
- *task + clue* based on features identified by the human-judges + *input*
*“This is a target audience identification task. Looking at the linguistic clues, identify for the target audience in which state of Brazil the has been constructed:
INPUT: XXXX
CLUES: interjections, second person pronoun, first person plural pronoun, second person verb agreement, first person plural verb agreement ”*

3.3. Data Wrangling

Data from the classification tasks was submitted to three stages of processing: 1) **Data cleaning**: city names were removed, retaining only state names; 2) **Standardization**: null answers were standardized as “não identificado,” while errors like “Sudoeste” and “Região Amazônica” were retained; and 3) **Data structuring**: all classification data was categorized and stored in .csv files.

The models generated a total of 108 texts, 4 for each Brazilian state (1 per model) and 216 classifications were carried out, 108 for each type of prompt (*task* and *task + clue*). The generated texts were also classified by three human experts in syntactic dialectal variation, with the task of identifying the target state of the text. Measurement of the classifications (both of LLMs and humans) were carried out using Fleiss' Kappa [Fleiss 1971].

4. Results

4.1. Exploratory analysis of target-profiles

The exploratory analysis revealed that Sabiá-2 [Sales Almeida et al. 2024], a Brazilian Portuguese-trained LLM from Maritalk, showed no dialectal variability in its responses.

- Sabiá-2 generation example:

“Oi, [NOME DO LUGAR] Você já conhece a nossa loja de camisetas sustentáveis? Se ainda não, prepare-se para descobrir um novo jeito de expressar o seu amor pela nossa terra e pelo nosso planeta! Na nossa produção, usamos algodão orgânico, que respeita o ciclo da vida e a terra que pisa o nosso pé, e tintas naturais, que não sujam as águas que correm pelo nosso estado. As estampas são uma homenagem ao nosso estado: cada ponto turístico, cada paisagem natural, cada aspecto da nossa rica cultura é representado com orgulho e carinho. Você vai encontrar desde o estilo básico até o mais ousado, mas sempre com aquele toque eco-friendly que só nós sabemos dar. E aí, vai deixar de fazer parte dessa revolução verde? Na nossa loja, além de encontrar camisetas que fazem a diferença, você vai ter a certeza de que cada compra fortalece a nossa comunidade e a nossa luta por um mundo mais sustentável. Venha nos visitar, escolha a sua favorita e mostre que em [NOME DO LUGAR], além de um povo acolhedor e de uma natureza exuberante, temos a consciência ambiental como uma das nossas maiores riquezas! Com amor e orgulho [NOME DO LUGAR], [NOME DA LOJA]”

In contrast, GPT 3.5, GPT-4o, and Gemini displayed sensitivity to dialectal differences, evident in variations of 2P pronouns (*tu* *você*), 1PP agreement (*nós vai nós vamos*), subjunctive/indicative morphology (*compre compra*), interjections (*uai, oxente, ôxe*), and discourse markers (*visse? tá ligado? é nóix, bora lá*).

- Gemini generation example:

“E aí, pessoal! Bem-vindos à [NOME DA LOJA]! Aqui no coração da Amazônia, nós tá fazendo moda de um jeito sustentável e cheio de orgulho da nossa terra. Nossas camisas são feitas com materiais reciclados e tinturas naturais, tudo bem cuidadoso pra respeitar o meio ambiente. O negócio é ficar estiloso e cuidar da nossa floresta ao mesmo tempo, visse? Aqui você encontra camisa que é a cara do [NOME DO LUGAR]: cheia de cor, vida e consciência. Vem com a gente nessa jornada de vestir a sustentabilidade com muito orgulho e amor pela natureza. Bora lá, junte-se a nós e venha conferir as novidades!”

- GPT 3.5 generation example:

“E aí, meu chegado?! Seja bem-vindo à [NOME DA LOJA], a loja onde a moda sustentável é raiz! Aqui, o negócio é respeitar a floresta e o nosso mundão, com camisetas que são a cara do [NOME DO LUGAR]? Nossas camisas são tudo de bom, feitas com material reciclado e tingidas do jeito mais consciente, porque a gente se importa com a Amazônia e com a natureza que tá no nosso sangue, né não? Então, se tu quer se vestir com estilo e ainda ajudar a proteger o planeta, vem com a gente! Aqui tu vai encontrar o visual perfeito e ainda fazer bonito com o meio ambiente. Quando quiser, dá um pulo aqui na loja que a porta tá sempre aberta pra ti, meu amigo! Fica à vontade, viu?”

- GPT-4o generation example:

“E aí, galera! Bem-vindos à [NOME DA LOJA]! Aqui é onde o estilo se encontra com o respeito pela natureza, do jeitinho que [NOME DO LUGAR] merece! Nós trabalhamos com material reciclado e tingimento responsável, tudo feito com carinho pras nossas terras [NOME DO LUGAR] e pro nosso planetão. É um compromisso sério, viu? Então, vem conferir nossas camisas que são um arraso, cheias de estilo e consciência ecológica. É moda que valoriza o que é nosso e cuida do futuro, firmeza? Bora junto nessa ideia de mostrar que em [NOME DO LUGAR], o verde tá na moda!”

These features align with the descriptive patterns identified by previous sociolinguistic studies [Abraçado and Martins 2015], reinforcing that some LLMs learn linguistic biases.

The generated responses were evaluated by three human judges specializing in syntactic dialectal variation. They identified the target state and provided reasoning for each input. Although their assessments highlighted similar dialectal clues as the exploratory analysis, the agreement was weak (inter-annotator: $\kappa = 0.31$; target: $\kappa = 0.13$). The next step was to have the LLMs perform the same task.

4.2. Target-profiles evaluation

LLMs were tasked with identifying the state of the text generated in the first step using two prompts: 1) *task + input* and 2) *task + clue + input*. The analysis flow is shown in Figure 3 and displays the Brazilian states, the LLMs used in both generation and classification, and the classifications performed with the two types of prompt. Although all prompts were in Brazilian Portuguese, Gemini responded in English.

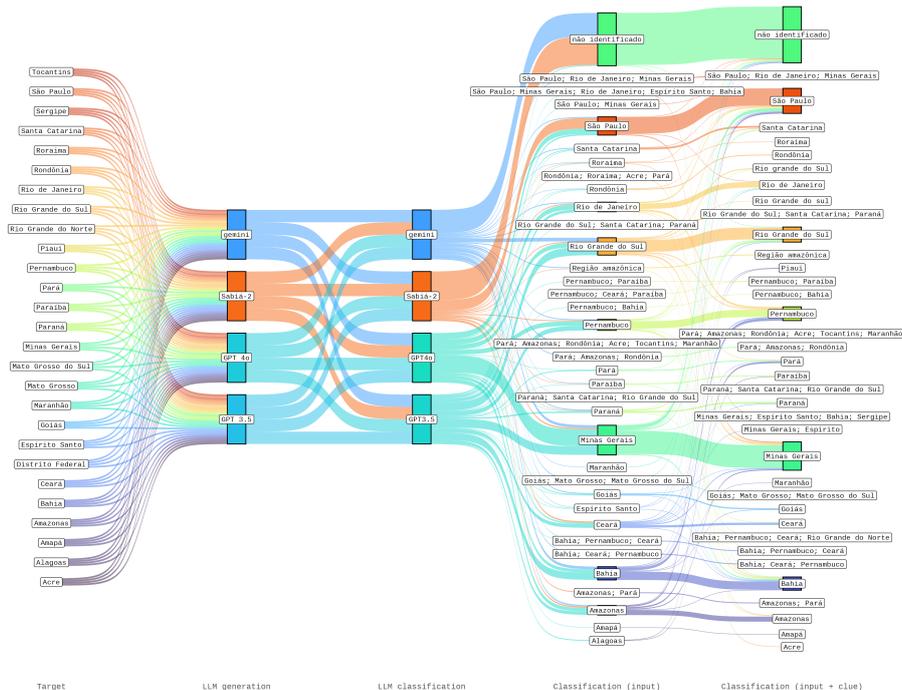


Figure 3. Analysis flowchart

In task 1, with weak agreement ($\kappa = 0.15$), differences emerged among LLMs: Sabiá-2 had the lowest agreement rate ($\kappa = 0.10$), while GPT-4o achieved the highest agreement, but still rather low ($\kappa = 0.21$). Identification rates varied as well: Sabiá-2 failed to identify a profile in 91% of cases, followed by Gemini at 88%, whereas GPT 3.5 and GPT-4o correctly identified the locations in over 20% of cases. Unexpectedly, in task 2, the non-identification rate increased (not significant), suggesting that the linguistic features identified by humans did not significantly influence the LLMs' classification. The only exception was GPT 3.5, which correctly identified one more location (an increase of 0.92%).

- This is an example of Sabiá-2's answer for task 1:

“Observando as marcas linguísticas e o conteúdo da propaganda, podemos determinar que o público-alvo é de um estado brasileiro com características culturais e ambientais específicas, que valoriza a sustentabilidade e tem um modo de falar próprio. [...] Considerando todas essas informações, é possível inferir que o público-alvo da propaganda é de um estado do Brasil com características culturais e ambientais específicas, que valoriza a sustentabilidade e tem um modo de falar próprio. Sem uma especificação clara do [NOME DA LOJA] ou do [NOME DO LUGAR], não é possível determinar exatamente qual estado é o público-alvo, mas as pistas sugerem um local com características semelhantes à região nordeste, dada a informalidade e o uso de gírias, ou talvez um estado do sul ou sudeste que tenha uma comunidade engajada em questões ambientais e use redes sociais como parte de sua estratégia de marketing.”

The input lacks specific geographic identification; it could be attributed to the “Nordeste” region due to informal style and slang, or to the “Sul” or “Sudeste” regions, presumed to have higher engagement in social issues. GPT 3.5, while imprecise in geographic localization, provides a more detailed identification of grammatical features even without additional clues:

“Com base nas marcas linguísticas presentes na propaganda do INPUT:

- **Uso de “aí” e “tá ligado”:** *Essas expressões são comuns em diversas regiões do Brasil, mas especialmente no contexto da propaganda, elas podem indicar uma influência mais forte de dialetos urbanos, como os encontrados em estados do Sudeste e Centro-Oeste.*
- **Realização do pronomes de primeira pessoa do plural:** *“Nóis trabalha firme...”, indicando um dialeto regional característico, comum em estados como São Paulo e Minas Gerais, por exemplo.*
- **Expressões regionais como “mais da hora”:** *Esta expressão é típica de regiões como o Sudeste e o Sul do Brasil.*
- **Referência ao compromisso com o planeta:** *A preocupação com sustentabilidade e meio ambiente é um tema geralmente valorizado em todo o Brasil, mas a forma como é apresentada na propaganda pode refletir uma sensibilidade urbana mais específica.*

*Com base nessas características linguísticas, a propaganda do INPUT parece ser direcionada principalmente para um público-alvo urbano, possivelmente das regiões **Sudeste ou Centro-Oeste do Brasil**, onde há uma mistura de expressões informais e um foco crescente em sustentabilidade.”*

GPT-4o provides the most accurate identification of the state intended by the input. The subsequent answer relates to the same input as those given by GPT 3.5 and Sabiá-2:

“Observando as marcas linguísticas presentes na propaganda, podemos identificar o público-alvo provável com base nos seguintes elementos:

- **Interjeições e expressões coloquiais:** "E aí", "tá ligado?", "nóis trabalha", "da hora", "bora", "sim senhor", "tá plantado".
- **Realização do pronome de segunda pessoa:** "você" não é explicitamente utilizado, mas a forma de tratamento é informal e coloquial.
- **Realização do pronome de primeira pessoa do plural:** "nóis" em vez de "nós".
- **Concordância verbal com segunda pessoa:** Não é explicitamente mostrada.
- **Concordância verbal com primeira pessoa do plural:** "nóis trabalha" em vez de "nós trabalhamos".

Essas marcas linguísticas são típicas do estado de São Paulo, especialmente da capital e regiões metropolitanas. O uso de "tá ligado?", "da hora" e "nóis" são característicos do dialeto paulista, que muitas vezes é associado à cultura urbana paulistana. Portanto, a propaganda do INPUT parece ser construída para o público-alvo do estado de São Paulo."

GPT-4o, considering all the grammatical features, determines that the input is intended for São Paulo, providing a more precise answer than GPT 3.5. Like humans, GPT-4o integrates linguistic clues and covariation, as illustrated in the following example:

"A análise das marcas linguísticas presentes no texto da propaganda do INPUT sugere que o público-alvo é o estado de Pernambuco. Aqui estão as razões:

1. **Interjeições e expressões regionais:** A expressão "visse?" é tipicamente pernambucana e nordestina.
2. **Realização do pronome de segunda pessoa:** A forma "você" é usada, que é comum em várias regiões do Brasil.
3. **Realização do pronome de primeira pessoa do plural:** "Nóis" em vez de "nós" indica uma variação regional.
4. **Concordância verbal com segunda pessoa:** "deixar você", "ajudar nosso planeta".
5. **Concordância verbal com primeira pessoa do plural:** "Nóis trabalha", uma forma de falar mais informal e regional.

Outras expressões como "galera", "arretadas de boas" e "bora" reforçam ainda mais o tom informal e típico da linguagem usada no estado de Pernambuco. Portanto, a propaganda do INPUT foi construída para o público-alvo de Pernambuco, usando elementos linguísticos característicos da região para criar uma conexão mais forte com os consumidores locais."

The feature associated with urbanity in the previous answer is seen as a regional trait in this case, leading to the attribution of the input to Pernambuco. This variation can be explained by indexical fields [Eckert 2008]. For instance, *nós vai* is considered an informal feature for Pernambuco but an indicator of urban culture for São Paulo. This pattern aligns with societal studies of Brazilian Portuguese [Freitag et al. 2015].

In task 1), the most common states identified were Minas Gerais (9% for GPT-4o and 4% for GPT 3.5), São Paulo (5% for Sabiá-2), and Rio Grande do Sul (2% for Gemini) (Figure 4). In task 2), results were similar except GPT 3.5, which identified Pernambuco (Figure 5). Only GPT 3.5 and GPT-4o provided explicit analyses of clues, such as *tu* being common in Southern Brazil or *nóis vai* indicating informality or urban features. These findings highlight the sociolinguistic fine-tuning of LLMs or their *language regard* [Preston 2010].

5. Discussion

This study investigated dialectal sensitivity in LLMs by assessing their responses to tasks aimed at identifying dialectal features in Brazilian Portuguese.

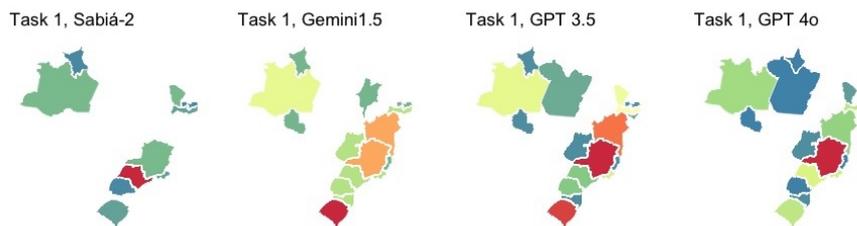


Figure 4. Geographical distribution of profile identification in task 1 *prompt: task + input*

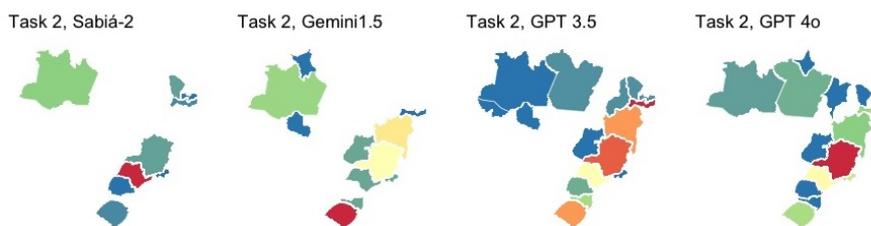


Figure 5. Geographical distribution of profile identification in task 2 *prompt: task + clue + input*

The findings revealed that Sabiá-2, a chatbot from Maritalk trained in Brazilian Portuguese, lacked dialectal sensitivity, showing no variability in responses. In contrast, GPT 3.5, GPT-4o, and Gemini demonstrated sensitivity to dialectal differences, evident in variations in pronoun usage, verb agreement, and other linguistic markers, aligning with sociolinguistic patterns. The agreement among human evaluators on dialectal features was weak ($\kappa = 0.31$), suggesting either inconsistent identification or a low number of judges. Among LLMs, Sabiá-2 had the lowest agreement rate ($\kappa = 0.08$), while GPT-4o showed the best agreement among the classifications ($\kappa = 0.21$). Notably, Gemini responded in English despite prompts being in Portuguese.

Incorporating specific linguistic features into prompts did not notably improve the LLMs' ability to identify the state, indicating that these features may not significantly affect classification. In task 1), GPT-4o and GPT 3.5 often identified Minas Gerais, while São Paulo and Rio Grande do Sul were identified by Sabiá-2 and Gemini1.5, respectively. In task 2), GPT 3.5 shifted its identification to Pernambuco. Only GPT 3.5 and GPT-4o provided explicit justifications based on dialectal clues, indicating some understanding of regional features.

These results show that while LLMs can detect dialectal variation, their ability to pinpoint specific regional profiles is inconsistent. The use of human-identified linguistic clues does not significantly enhance classification accuracy. Understanding how LLMs handle language varieties can help sociolinguistics explore human processing of linguistic variation and contribute to advancing linguistic justice and equitable NLP technologies [Baugh 2018, Wolfram and Eisenhauer 2019, Nee et al. 2021, Nee et al. 2022, Liu et al. 2023].

References

- Abraçado, J. and Martins, M. A. (2015). *Mapeamento sociolinguístico do português brasileiro*. Editora Contexto.
- Baugh, J. (2018). *Linguistics in pursuit of justice*. Cambridge University Press.
- Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. (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*, pages 610–623.
- Dacon, J. (2022). Towards a deep multi-layered dialectal language analysis: A case study of african-american english. *arXiv preprint arXiv:2206.08978*.
- Dacon, J., Liu, H., and Tang, J. (2022). Evaluating and mitigating inherent linguistic bias of african american english through inference. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1442–1454.
- Dacon, J. and Tang, J. (2021). What truly matters? using linguistic cues for analyzing the# blacklivesmatter movement and its counter protests: 2013 to 2020. *arXiv preprint arXiv:2109.12192*.
- Danziger, S. (2018). Where intelligence lies: Externalist and sociolinguistic perspectives on the turing test and ai. In *Philosophy and Theory of Artificial Intelligence 2017*, pages 158–174. Springer.
- Eckert, P. (2008). Variation and the indexical field. *Journal of sociolinguistics*, 12(4):453–476.
- Fleisig, E., Smith, G., Bossi, M., Rustagi, I., Yin, X., and Klein, D. (2024). Linguistic bias in chatgpt: Language models reinforce dialect discrimination. *arXiv preprint arXiv:2406.08818*.
- Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.
- Freitag, R. M. K. (2016). Sociolinguística no/do brasil. *Cadernos de Estudos Linguísticos*, 58(3):445–460.
- Freitag, R. M. K. (2021). Preconceito linguístico para humanizar as máquinas. *Cadernos de Linguística*, 2(4):e495–e495.
- Freitag, R. M. K., Severo, C. G., Rost-Snichelotto, C. A., and Tavares, M. A. (2015). Como o brasileiro acha que fala? desaios e propostas para a caracterização do” português brasileiro”. *Signo y seña*, (28):65–87.
- Freitag, R. M. K., Severo, C. G., Rost-Snichelotto, C. A., and Tavares, M. A. (2016). Como os brasileiros acham que falam? percepções sociolinguísticas de universitários do sul e do nordeste. *Todas as Letras-Revista de Língua e Literatura*, 18(2).
- Grieve, J., Bartl, S., Fuoli, M., Grafmiller, J., Huang, W., Jawerbaum, A., Murakami, A., Perlman, M., Roemling, D., and Winter, B. (2024). The sociolinguistic foundations of language modeling. *arXiv preprint arXiv:2407.09241*.
- Liu, Y., Held, W., and Yang, D. (2023). Dada: Dialect adaptation via dynamic aggregation of linguistic rules. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13776–13793.

- Mengesha, Z., Heldreth, C., Lahav, M., Sublewski, J., and Tuennerman, E. (2021). “i don’t think these devices are very culturally sensitive.”—impact of automated speech recognition errors on african americans. *Frontiers in Artificial Intelligence*, 4:725911.
- Nee, J., Macfarlane Smith, G., Sheares, A., and Rustagi, I. (2021). Advancing social justice through linguistic justice: Strategies for building equity fluent nlp technology. In *Proceedings of the 1st ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, pages 1–9.
- Nee, J., Smith, G. M., Sheares, A., and Rustagi, I. (2022). Linguistic justice as a framework for designing, developing, and managing natural language processing tools. *Big Data & Society*, 9(1):20539517221090930.
- Preston, D. R. (2010). Language, people, salience, space: perceptual dialectology and language regard. *Dialectologia: revista electrònica*, pages 87–131.
- Roncarati, C., Abraçado, J., and Heye, J. B. (2003). *Português brasileiro: contato lingüístico, heterogeneidade e história*, volume 2. 7 Letras.
- Sales Almeida, T., Abonizio, H., Nogueira, R., and Pires, R. (2024). Sabiá-2: A new generation of portuguese large language models. *arXiv e-prints*, pages arXiv–2403.
- Shrawgi, H., Rath, P., Singhal, T., and Dandapat, S. (2024). Uncovering stereotypes in large language models: A task complexity-based approach. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1841–1857.
- Wolfram, W. and Eisenhauer, K. (2019). Implicit sociolinguistic bias and social justice. In *The Routledge Companion to the Work of John R. Rickford*, pages 269–280. Routledge.

Mini-glossário do Tucumã do Pará no Município de Acará: olhares, significados e cultura da Amazônia

Eliene da S. Alves¹, Brayna C. dos S. Cardoso²

¹Discente do Programa de Pós-Graduação em Cidades: Territórios, Identidades e Educação – PPGCITE - Universidade Federal do Pará (UFPA)
Abaetetuba – PA – Brasil

² Docente do Instituto de Letras e Comunicação e do Programa de Pós-Graduação em Cidades: Territórios, Identidades e Educação – PPGCITE - Universidade Federal do Pará (UFPA)
Belém/Abaetetuba – PA – Brasil

eliene.alves@abaetetuba.ufpa.br, braynacardoso@ufpa.br, ronaldosousa@ufpa.br

Abstract. *This article aims to analyze the possible socioterminological meanings linked to Tucumã do Pará and its processes in the experiences of subjects in the municipality of Acará-PA. It is a qualitative and descriptive research on the process of structuring and preparing the mini-glossary. Which was theoretically based on the studies of Brandão (2007) addressing knowledge; Cimerys (2005) on Tucumã do Pará; and Faulstich (1995) in producing the glossary. With this, the mini-glossary of Tucumã do Pará, with its variations and meanings, became a favorable tool for the terminological knowledge of socio-professional activity in the Amazon of Pará and in the dissemination of the bioeconomy of the craft that it provides.*

Resumo. *Este artigo se propõe a analisar os possíveis significados socioterminológicos vinculados ao Tucumã do Pará e seus processos na vivência de sujeitos do município de Acará-PA. É uma pesquisa qualitativa e descritiva sobre o percurso de estruturação e elaboração do mini-glossário. O qual embasou-se teoricamente nos estudos de Brandão (2007) abordando os saberes; Cymerys (2005) sobre o Tucumã do Pará; e Faulstich (1995) na produção do glossário. Com isso, o mini-glossário do Tucumã do Pará com suas variações e significados, tornou-se uma ferramenta favorável para o conhecimento terminológico da atividade socioprofissional na Amazônia paraense e na difusão da bioeconomia do ofício que ele proporciona.*

1. Introdução

As vozes da Amazônia ecoam no mundo, seja pela vívida autenticidade que esta possui com suas exuberantes faunas e floras ou pela importância nítida que tem para o planeta. Com isso, quem vive na Amazônia está intrinsecamente acostumado a olhar os céus, as matas, os rios e os animais como aspectos favoráveis a um cotidiano onde a vida move-se de maneira mais tranquila nas regiões com maior quantidade de florestas, quanto mais a fundo entra-se na Amazônia paraense, mais suave se percebem os dias, sejam pelos

coloridos vibrantes de algumas plantas ou pelas nuances de tonalidades de verde das árvores, tudo se transforma e se adapta ao longo dos diversos lugares amazônicos.

No município de Acará, onde as florestas e as águas percorrem grande parte do território percebe-se essa conexão atrativa com e pela natureza nos mais variados lugares, a cidade à margem direita do rio acará, apresenta uma mesclagem de natureza/cidade, rural/urbano, onde a interação com o rio, os córregos e igarapés são presentes no cotidiano dos sujeitos.

Ao fazer o movimento de trânsito entre as idas e vindas do campo para a cidade, compreendeu-se a interação ativa do Tucumã do Pará, *Astrocaryum vulgare* Mart., pelas vozes dos sujeitos com vínculos socioterminológicos específicos durante a manipulação do tucumã. Vozes que ecoavam significados e construções características que possuem sentidos próprios para a vivência do tucumã.

Dessa forma, esta pesquisa tem como objetivo principal analisar os possíveis significados socioterminológicos que se vinculam ao Tucumã do Pará e seus processos na vivência de sujeitos do município de Acará-PA. Para tanto, tornou-se necessário identificar as variações de significados para estruturas que compõe o Tucumã do Pará; elaborar um mini-glossário socioterminológico a respeito do Tucumã do Pará e seus usos; e apresentar os significados sobre o Tucumã do Pará como fonte de conhecimento terminológico da atividade socioprofissional desempenhada na Amazônia Paraense e na difusão da bioeconomia que este ofício proporciona.

Para tanto, ressalta-se a importância no processo de execução da construção do mini-glossário por meio de aporte tecnológico computacional, com uso do *canva* como instrumento facilitador para a elaboração do design gráfico do mini-glossário.

Como forma de esboçar um panorama geral do que aqui é discutido, este artigo apresenta a seção introdutória; o referencial teórico, com as vozes científicas a respeito de ações lexicais, socioterminológicas, o mini-glossário e o Tucumã do Pará; assim como a metodologia; os resultados e discussões; e as considerações finais.

2. Referencial Teórico

Quando se abordam diálogos a respeito dos significados de palavras na Amazônia Paraense as variações são nítidas, dependendo das regiões do estado, das áreas urbanas ou rurais, das idades, entre outros, esses fatores são predominantes devido as variáveis lexicais que há na Amazônia Brasileira por ser um lugar com diferentes culturas e aspectos socioculturais atrelados as vivências dos sujeitos em relação com a natureza.

Portanto, para compreender essas ações lexicais gerais e específicas, faz-se necessário observar os conceitos a respeito de lexicologia e lexicografia; terminologia; socioterminologia; e glossário. Consequentemente, a lexicologia estuda os aspectos teóricos, e a lexicografia estuda a sistematização e categorização para análise dos significados das palavras, fazendo referência a dicionários e glossários [Sá, 2020; Lisboa, 2015; Barbosa, 1992]. Para tanto, segundo Barbosa (1992) a lexicografia:

“(...) se define como uma tecnologia de tratamento daquela, de compilação, classificação, análise e processamento, de que resulta, por exemplo, a produção de dicionários, vocabulários técnicos científicos, vocabulários especializados e congêneres.” [Barbosa, 1992, p. 4]

Por conseguinte, a lexicografia aponta-se como fundamental para essa proposta, uma vez que se propõe a dialogar sobre a produção de um mini-glossário. Outro conceito de relevância é a terminologia que segundo Faulstich (2001), apresentam-se aspectos gerais e variacionistas, em consonância com as diversidades que a língua apresenta. Assim, a terminologia proporciona uma ação atuante na formulação dos significados.

Para tanto, com a intensa influência social a respeito das terminologias desenvolveu-se a socioterminologia, que segundo Faulstich (1995) se apresenta como uma proposta de interação entre os aspectos sociais e a terminologia. Com isso, a socioterminologia “(...) como prática do trabalho terminológico, fundamenta-se na análise de circulação do termo no funcionamento da linguagem.”, logo, atuando “(...) como disciplina descritiva, estuda o termo sob a perspectiva linguística na interação social.” [Faulstich, 1995, p. 2] aportando-se nos princípios da sociolinguística e da etnografia para proceder a pesquisa socioterminológica.

Destarte, a pesquisa socioterminológica está atrelada a estes dois conceitos fundamentais, segundo Faulstich (1995), a sociolinguística que analisa as relações sociais com vínculos linguísticos, e a etnografia que se aproxima ativamente das interações sociais da vivência no cotidiano dos sujeitos.

Logo, Faulstich (1995) descreve que a socioterminologia utiliza procedimentos provenientes da etnografia integrados com suas definições e aspectos sociais, sendo necessário observar fatores como: as características em que a terminologia é gerada; as características dos sujeitos; a competência e os usos linguísticos; o uso de instrumentos tecnológicos para caracterizar e registrar a variação linguística na terminologia.

Para tal propósito, o glossário define-se como “Pequeno vocabulário, ou relação de palavras, em que se explica o significado das mesmas para ajudar o leitor na compreensão do texto que lê.” [Biderman, 1984, p. 139] constituindo uma linguagem técnica.

Deste modo, conforme Faulstich (1995) na elaboração de um glossário faz-se necessário analisar:

a) Repertório que define termos de uma área científica ou técnica, dispostos em ordem alfabética, podendo apresentar ou não remissivas.

b) Repertório em que os termos, normalmente de uma área, são apresentados em ordem sistemática, acompanhados de informação gramatical, definição, remissivas podendo apresentar ou não contexto de ocorrência.

Nota: os glossários em ordem alfabética e os em ordem sistemática podem também conter sinonímia, variante(s) e equivalente(s).

c) Repertório em que os termos são apresentados em ordem alfabética ou em ordem sistemática seguidos de informação gramatical e do contexto de ocorrência.

Nota: este tipo de glossário é útil para tradutores e intérpretes; elabora-se, normalmente, a partir de bases textuais informatizadas. [Faulstich, 1995, p. 6]

Com isso, na produção do material às oralidades sociais, verificações teóricas, organizações de termos e dos significados são essenciais para obter resultados positivos com tais materiais. Deste modo, as relações conceituais são naturalmente fundamentais para a compreensão das palavras utilizadas de acordo com as ações realizadas e/ou contexto no qual está inserido determinada palavra.

Então, a partir dessas compreensões teóricas sobre ações lexicais sendo possível o entendimento sobre a lexicografia e as nuances para compor um glossário socioterminológico, segue-se para a compreensão a respeito dos conceitos sistematizados sobre a palmeira do Tucumã do Pará, que norteia a edição do mini-glossário na Amazônia Paraense.

O Tucumã do Pará denominado de *Astrocaryum vulgare* Mart. é uma palmeira típica na Região da Amazônia Paraense, pode ser vista em diversos lugares por se adaptar a solos variados [Shanley; Medina, 2005], segundo Ribeiro *et al* esta palmeira “Cresce próximo de rios, em áreas não cobertas com água, em terra firme, cobertura vegetal baixa e em campo limpo.” (2014, p. 2), “Essa palmeira é considerada uma planta pioneira e invasora de pastos, mas também é encontrada em capoeiras e florestas. Desenvolve-se bem em solos pobres de terra firme.” [Cymersys, 2005, p. 209], sendo também resistente às queimadas.

Outrora, seu fruto também serve de alimento para alguns animais na fauna amazônica [Cymersys, 2005], porém o *A. vulgare* possui diversas utilidades para os humanos, o fruto pode ser consumido *in natura* ou processado, em alimento de animais como o porco e a galinha; caroço na confecção de artesanatos; óleo em preparos alimentícios, na produção de cosméticos; palha na produção de telhados e de artesanatos; espinhos na confecção de artesanatos; e o tronco no uso construções de pequeno porte [Shanley; Medina, 2005; De Menezes *et al*, 2012].

A relação com *A. vulgare* torna-se cultural, econômica, ancestral e simbólica por fazer parte de alguns espaços sociais de convívio na região da Amazônia Paraense [Medeiros, 2012; Silva, 2019; Silva, 2021]. Com isso, a seguir apresenta-se o percurso metodológico para a elaboração do mini-glossário socioterminológico do tucumã do Pará.

3. Metodologia

O trabalho foi organizado com base na ideia central de compreensão a respeito dos possíveis significados socioterminológicos que se vinculam ao Tucumã do Pará e seus processos na vivência de sujeitos do município de Acará-PA. Para tanto, este estudo apresenta-se como proposta descritiva por intermédio da organização, estruturação e elaboração do mini-glossário do Tucumã do Pará.

Quanto ao critério qualitativo, este foi realizado por meio das observações atreladas as vivências de pessoas que possuíam vínculos de uso do Tucumã do Pará, com participações a partir das oralidades e simbologias [Minayo *et. al*, 2009], a coleta de dados para a identificação das socioterminologias foi realizada durante as entrevistas, assim como, pelos diálogos no percurso realizado nas visitas de campo denominadas de turnê guiada, que consiste em realizar a visita na propriedade que contém o tucumã e acompanhar os sujeitos desde a residência até a área onde possui tucumã, sendo orientados por eles sobre os usos ativos do tucumã na propriedade [Albuquerque *et al.*, 2010], nesse trajeto utilizou-se o caderno de campo para realizar as anotações das socioterminologias e os significados para os participantes da pesquisa, e posteriormente, foram observadas as palavras com maior repetição entre eles como critério de seleção das palavras para a produção dos significados e elaboração do mini-glossário.

Para tanto, o *design* gráfico do mini-glossário do tucumã do Pará ocorreu por meio do uso da plataforma *canva* (versão pro), a partir do auxílio de instrumentos editáveis em uma folha em branco, com construções criativas, fruto das cognições das autoras.

Neste sentido, durante os diálogos anotou-se as palavras que possuíam significados socioterminológicos pelas relações destes com o tucumã e de maior frequência nos encontros e na turnê guiada, no diálogo informal e nas entrevistas.

Logo, as palavras foram categorizadas por meio da subdivisão em Partes do Tucumã, objetivando apontar as descrições da composição de uma palmeira do tucumã com as nomenclaturas principais descritas pelos entrevistados; Usos da polpa do Tucumã, com as diversas possibilidades de manipulações da polpa para consumo; e Usos do caroço do Tucumã, no processo de produção de biojóias.

Com isso, após a seleção e divisão das palavras em categorias, foram realizados os apontamentos de informações gramaticais como: Substantivo; Sílabas Tônicas; Plural; Variação; e Dados científicos. Sendo necessário esta organização das informações gramaticais em uma construção de glossário [Faustich, 1995], assim como os dados científicos sobre a morfologia do tucumã.

Na sequência, foram organizados os significados, onde se descreve as características, funcionalidades e repercussão do tucumã do Pará e, isso torna-se fundamental para a compreensão do leitor, a respeito da palavra apontada no glossário [Biderman, 1984].

Além disso, foram acrescentadas fotografias das partes a serem citadas no mini-glossário de tucumã do Pará e um exemplo da palavra em uma frase, coletadas nos diálogos com os entrevistados. Consequentemente, para facilitar a compreensão do significado dentro do mini-glossário as imagens foram adicionadas como instrumento simbólico de identificação nas estruturas, onde a fotografia tem como fundo uma pausa na imagem e proporciona vida e significados ao contexto que palavras podem reduzir o alcance da expressão [Brandão, 2004], assim como, a inclusão da palavra inserida em uma frase no contexto utilizado pelos sujeitos da pesquisa com raízes nas vivências.

Desse modo, foi possível estruturar o mini-glossário com descrições separadas pelas subdivisões de estruturas e usos a respeito do tucumã, e em cada divisão separou-se as palavras por ordem alfabética, conforme as indicações de Faustich (1995).

Logo, a elaboração do mini-glossário a respeito dos significados sobre o tucumã e suas partes, apresenta-se como proposta para facilitar a compreensão a respeito das estruturas do tucumã do Pará, uma palmeira típica da região amazônica, porém com observações e usos bem específicos para pessoas que possuem vínculos ancestrais ativos com o tucumã. Com isso, no tópico a seguir será apresentado os olhares e saberes para a elaboração do mini-glossário do tucumã do Pará.

4. Resultados e discussão

Entre o olhar e o saber estão a inquietude de quem pesquisa, quem observa e aprende com as práticas e as oralidades dos que cotidianamente vivenciam as experiências com esse fruto denominado tucumã.

Destarte, percorrem-se as aprendizagens que fluem na pesquisa sobre o Tucumã do Pará pelas intertrocas de saberes [Brandão, 2007], *saber* de um povo que vive e revive essa prática como algo que transcende sua vida, se mescla a vida de seus ancestrais que na prática e oralidade lhe ensinaram a observar a natureza e a realizar as manipulações necessárias para fazerem uso dos frutos amazônicos, como o Tucumã do Pará.

E com um *olhar* inquieto, por um prisma de curiosidade, mas com a maturidade de pesquisa, voz respeitosa e diálogos breves, foram-se tecendo os questionamentos dos significados e como uma luz reluzente tudo fez sentido a respeito do tucumã. A exemplo, a palavra *facão*, que ao primeiro som da palavra remeteu a memória, a descrição pelo significado do dicionário, como ferramenta para uso, e segundo o Dicionário *Online* de Português a palavra *facão* significa “Grande faca”, contudo, na exposição dos sujeitos, *facão*, no tucumã do Pará, é a estrutura do tucumanzeiro que protege a flor durante a fase de inflorescência, obtendo um conceito socioterminológico específico de acordo com a vivência dos sujeitos, assim como outras palavras descritas no mini-glossário do tucumã do Pará.

Deste modo, foram selecionados vinte termos para compor o mini-glossário do tucumã do Pará, as terminologias a respeito das partes dos tucumã foram: *Barca de tucumã*, que diz respeito a estrutura do tucumanzeiro que protege o fruto até soltar do cacho, cientificamente pode ser denominada de espata, e sua variante socioterminológica é a canoa de tucumã; *Bicho de tucumã* que é uma larva que consome a parte interna da semente do caroço do tucumã, cientificamente denominado de *Speciomerus ruficornis* Germar; *Broto do tucumã* que é o processo de germinação (processo de crescimento de uma nova planta a partir da semente) para o desenvolvimento do tucumanzeiro, e cientificamente denomina-se semente em fase de germinação.

Assim como, o *Caroço de tucumã*, é a parte do fruto do tucumã onde se encontra o coquinho, cientificamente chama-se endocarpo lenhoso; *Espinho de tucumã*, é a parte pontiaguda que está presente nas diversas partes do tucumanzeiro, cientificamente esta estrutura pontiaguda denomina-se espina; *Facão de tucumã*, é uma estrutura de proteção da flor de tucumã, cientificamente também denomina-se espata; *Flor de tucumã*, são as flores do tucumanzeiro, cientificamente são inflorescências interfolares, ramificadas e eretas; *Palha de tucumã*, são os filamentos lisos e espinhosos que compõem o tucumanzeiro, possuem variantes sendo denominadas de folhas, e cientificamente também denominam-se folhas pinadas, reduplicadas e ascendentes com espinhos na nervura central; *Tucumã*, é o fruto do tucumanzeiro, cientificamente divide-se em epicarpo liso, mesocarpo fibroso, endocarpo lenhoso e endocarpo; *Tucumanzeiro*, é uma palmeira nativa da região amazônica que pode ser encontrada em áreas de terra firme, cientificamente denomina-se *Astrocayum vulgare* Mart.

Neste sentido, as terminologias selecionadas referente aos produtos feitos do tucumã do Pará foram por meio do resultado do processo de extração das duas partes externas do fruto popularmente chamada de casca e massa, cientificamente corresponde com a extração do epicarpo e mesocarpo, resultando no *Chopp de tucumã*, líquido do tucumã com adição de açúcar e processamento, este líquido congelado torna-se no chopp; *Creme de tucumã*, próximo ao processamento para elaboração do chopp, contudo adiciona leite condensado e creme de leite, e leva ao congelador, ideal para servir gelado; *Polpa de tucumã*, é o líquido processado que pode ser consumido in natura e/ou processado para consumo, como descrito neste; *Tucupi*, é produzido a partir do líquido proveniente da polpa com adição de água e fervura até possuir a consistência de tucupi; *Vinho de tucupi*, sua produção provém do líquido consistente da polpa para consumo com alimentos.

Para tanto, também foram selecionadas terminologias a respeito dos produtos feitos do caroço do tucumã, que tem como nome científico endocarpo lenhoso no qual é

produzido o *Anel de tucumã*, *Aliança de tucumã*, *Anel de tucumã com gravura* e o *Anel de tucumã com dedicatória*, que se diferenciam pelas características de composições e estruturas dos formatos, outra terminologia relevante que foi caracterizada foi o *Óleo do bicho de tucumã*, que é a extração por aquecimento da larva denominada *Speciomerus ruficornis* Germar que cresce dentro do caroço de tucumã.

Em suma, tornou-se possível realizar a elaboração deste mini-glossário, a partir da plataforma computacional *canva*, onde foram realizadas as organizações e gerência para a estruturação, sendo assim, fundamental no processo de construção do mini-glossário. Sua visibilidade favorece também no entendimento das variações e significações lexicais de acordo com as estruturas socioterminológicas que estavam sendo utilizada nas nuances que os colaboradores da pesquisa declaravam, pois para cada palavra descrita há um entendimento baseado nas relações socioculturais destes sujeitos.

Portanto, o *olhar* de inquietude ressignificou-se em novas traduções para o vivenciado com vínculo ao *saber* sobre o tucumã, onde as interações com os sujeitos transformaram-se em intertrocas de saberes [Brandão, 2007] para aprender a *olhar* para além do que se observa em uma planta como o tucumã, que no *saber* dos sujeitos está preenchido de história, memória e ancestralidade, e principalmente que a relação com natureza é vívida e movente aos que nela se fazem vida e se mesclam a esse lugar denominado Amazônia Paraense. A seguir, são apresentadas as considerações finais.

5. Considerações finais

A interação com a natureza amazônica se aprende no *olhar* e no *saber*, ao observar a prática e/ou a oralidade dos sujeitos, ao dialogar com quem ativamente está mesclando-se com a natureza, manipulando seus frutos e/ou consumindo de produtos provenientes desses lugares.

Para tanto, isso também ocorre com quem faz uso do Tucumã do Pará, onde este faz-se presente ativamente na vida dos sujeitos, realizando uma interconexão dos saberes sobre a palmeira e seus usos, com terminologias tão específicas que se tornam desconhecidas, aos que não fazem uso no mesmo contexto, porém ao interagir aprendem e compreendem fluentemente a conexão das socioterminologias pelos contextos.

Com isso, foi possível observar a importância do uso tecnológico no processo de construção do mini-glossário, onde a partir da edição gráfica no aplicativo computacional foram realizadas as organizações que possibilitam uma maior visualização e compreensão a respeito das socioterminologias vinculadas ao Tucumã do Pará.

Logo, analisar os significados socioterminológicos que se vinculam ao Tucumã do Pará e seus processos na vivência de sujeitos do município de Acará-PA, tornou-se elemento de importância social, científica e ambiental.

Neste sentido, a compreensão dos significados socioterminológicos a respeito do tucumã do Pará favorece o entendimento dos sujeitos e fomenta a valorização da identidade dos sujeitos que estão nesses lugares, bem como fortalece o reconhecimento dos sujeitos, pelos costumes ancestrais, que por vezes são desvalorizados pelas gerações seguintes.

Deste modo, o entendimento sobre as nomenclaturas e vínculos socioterminológicos do tucumã do Pará impulsiona inquietudes sobre mais investigações

desta palmeira e seus vínculos sociais e culturais que, em relação a outras palmeiras amazônicas, ainda possuem um número reduzido de pesquisas na área.

Além disso, favorece ambientalmente no fortalecimento para a proteção e cuidado com a palmeira do tucumã nos diversos espaços, proporcionando a permanência desta palmeira nativa na região amazônica, por meio da disseminação dessa realidade socioprofissional e proporcionando um avanço para a bioeconomia da cidade do Acará.

Referências

- Albuquerque, U. P; Lucena, R.F.P; Lins Neto, E. M. F. “Seleção dos participantes da pesquisa”. In: Albuquerque, U.P; Lucena, R. F. P; Cunha, L. V. F. C. (Org.). Métodos e técnicas na pesquisa etnobiológica e etnoecológica. Recife, PE: NUPEEA, 2010.
- Barbosa, Maria Aparecida. “Lexicologia, lexicografia, terminologia, terminografia, identidade científica, objeto, métodos, campos de atuação”. In: Anais, 1992.
- Biderman, Maria Tereza Camargo. “Glossário”. *ALFA: Revista de Linguística*, São Paulo, n. 28, 1984, p. 135-144.
- Brandão, Carlos Rodrigues. “Fotografar, documentar, dizer com a imagem”. In: *Cadernos de Antropologia e Imagem*, Rio de Janeiro, 18: 27-57, 2004.
- _____. Tempos e espaços nos mundos rurais do Brasil. *RURIS*: Campinas, v. 1, n. 1, 2007, p. 37-64.
- Cymerys, M. “Tucumã-do-Pará”. In: Shanley, P; Medina, G. Frutíferas e Plantas úteis na vida Amazônica. Belém: CIFOR, Imazon, 2005. p. 209-214.
- De Menezes, A. J. E. A.; Homma, A. K. O.; Oliveira, M. E. C.; De Matos, G. B. “Exploração do óleo de tucumã do pará (*Astrocaryum vulgare Mart.*) na mesorregião da ilha do Marajó-município de Soure-Pará”. In: Congresso Brasileiro de Recursos Genéticos, 2, 2012, Belém, PA. Anais... Brasília, DF: Sociedade Brasileira de Recursos Genéticos, 2012.
- Faustich, Enilde Leite de Jesus. “Aspectos de terminologia geral e terminologia variacionista”. *Tradterm*, v. 7, 2001, p.11-40.
- _____. “Socioterminologia: mais que um método de pesquisa, uma disciplina”. *Ciência da Informação*, v. 24, n. 3, 1995.
- “Facão”. In: DICIO, Dicionário Online de Português. Porto: 7Graus, 2020. Disponível em: <https://www.dicio.com.br/facao/>. Acessado em: 07 de julho de 2024.
- Lisboa, Josué Leonardo Santos de Souza. “Terminologia da piscicultura”. Orientador: Abdelhak Razky. 2015. 173 f. Dissertação (Mestrado) - Universidade Federal do Pará, Instituto de Letras e Comunicação, Belém, 2015. Programa de Pós-Graduação em Letras. Disponível em: <http://repositorio.ufpa.br/jspui/handle/2011/10170>. Acesso em: 25 de junho de 2024.
- Medeiros, Thais Helena. “Artesanias em palha de tucumã e memória: tecendo territorialidade e relações socioculturais”. *Somanlu: Revista de Estudos Amazônicos*, v. 12, n. 2, p. 151-173, 2012.
- Minayo, Maria Cecília de Souza; Deslandes, Suely Ferreira; Gomes, Romeu. Pesquisa Social: teoria, método e criatividade. Petrópolis, RJ: Editora Vozes, 28 ed., 2009.

- Ribeiro, L. L.; Lima, I.; Cunha, L.; Pacheco, E.; Silva, R. T. “Biometria dos frutos de tucumã (*Astrocaryum vulgare* Mart.) no município de Capitão Poço/PA”. Enciclopédia Biosfera, v. 10, n. 19, 2014.
- Sá, Edimilson José de. “Lexicografia e Geolinguística: um pequeno glossário de itens lexicais retirados de atlas linguísticos pernambucanos”. Revista do GELNE, Natal/RN, v. 22, n. 1, 2020. p.101-115.
- Shanley, Patricia; Medina, Gabriel. Frutíferas e plantas úteis na vida amazônica. Belém: CIFOR, Imazon, 2005. p. 209-214.
- Silva, Daniella Amor Cunha da. Potencialidade do tucumã (*Astrocaryum vulgare* Mart.) no município de Irituia-Pa: um novo produto para cooperativa d’Irituia. 2019. 43 f. Trabalho de Conclusão de Curso (Engenharia Agrônoma) – Universidade Federal Rural da Amazônia, Capitão Poço, 2019.
- Silva, Andrea Araújo da *et al.* Manejo, extração, uso e beneficiamento da palha do tucumã por mulheres da reserva extrativista Tapajós-Arapiuns, Pará, Brasil. 2021. Tese de Doutorado. Universidade Federal do Oeste do Pará.

Desambiguação de lema e atributos morfológicos na anotação do *corp*us Porttinari-base

Lucelene Lopes¹, Magali S. Duran¹, Thiago Alexandre Salgueiro Pardo¹

¹Instituto de Ciências Matemáticas e de Computação
Universidade de São Paulo, São Carlos/SP, Brazil

Abstract. *This paper reports the process of disambiguating lemmas and morphological features in a corpus of Portuguese annotated with Universal Dependencies tagset. We explain the strategies adopted to simplify and reduce the workload of annotators. These strategies contribute to improve the accuracy of linguistic annotation, which is fundamental for various Natural Language Processing tasks.*

Resumo. *Este artigo relata o processo de desambiguação de lemas e atributos morfológicos em um *corp*us do português anotado com os conjuntos de etiquetas da Universal Dependencies. Explicamos as estratégias adotadas para simplificar e diminuir o trabalho dos anotadores. Essas estratégias contribuem para aumentar a precisão da anotação linguística, a qual é fundamental para diversas tarefas de Processamento de Linguagem Natural.*

1. Introdução

Quanto mais interpretações humanas, lógicas e objetivas, estiverem registradas em um *corp*us sob forma de anotação, mais ricas as possibilidades de se utilizar esse *corp*us em tarefas de Processamento de Linguagem Natural (PLN). Uma anotação consistente é particularmente importante para construir um *corp*us de treino para modelos automáticos [Goldberg 2015, Gamba and Zeman 2023]. Cientes disso vários cientistas uniram seus esforços e vêm aperfeiçoando um modelo multilíngue de anotação chamado *Universal Dependencies* (UD) [Nivre et al. 2016, de Marneffe et al. 2021].

Entre as anotações previstas pela UD, estão a informação do lema, dos atributos morfológicos e da etiqueta morfossintática (*Universal Part-of-Speech* – UPOS) de cada token. Essas informações, muitas vezes, podem ser atribuídas automaticamente, utilizando-se um léxico computacional. No entanto, para tokens que apresentam mais de uma possibilidade de UPOS, lema e/ou atributos morfológicos, é preciso que os anotadores humanos decidam pelas informações mais apropriadas. Em PLN, essas múltiplas possibilidades de anotação são tratadas como casos de ambiguidade, embora, para um humano, o contexto quase sempre seja suficiente para determinar essas informações.

A desambiguação de UPOS em língua portuguesa, usando etiquetas da UD, já foi objeto de outros trabalhos [Lopes et al. 2023, Duran et al. 2022, Duran et al. 2021], porém, até onde é de nosso conhecimento, este é o primeiro trabalho a discutir especificamente a tarefa de desambiguação de lemas e atributos morfológicos em um *corp*us do português. Compartilhamos os métodos empregados para pré-categorizar automaticamente os tipos de ambiguidades, destacamos casos computacionalmente ambíguos do léxico do português, e apresentamos as quantidades de tokens e sentenças revisados e desambiguados no *corp*us Porttinari-base [Duran et al. 2023].

Organizamos o artigo como segue: na Seção 2, apresentamos informações sobre a abordagem UD, com vistas a fornecer conceitos básicos utilizados ao longo das discussões; na Seção 3, delineamos a metodologia adotada; na Seção 4, descrevemos em detalhe o processo de desambiguação realizado; na Seção 5, resumimos as contribuições deste trabalho e descrevemos as lições aprendidas.

2. Contextualização

A abordagem UD tem sido utilizada para anotar diversos corpú de várias línguas, incluindo os corpú de português, como o Bosque [Rademaker et al. 2017], o Petrogold [Souza et al. 2021], o Portinari [Duran et al. 2023] e o Cintil [Branco et al. 2022]. O formato de arquivo usado para a anotação UD é o formato CoNLL-U, que contém 10 colunas [Universal Dependencies 2023]. Cada coluna do CoNLL-U tem um nome e as colunas que citaremos neste artigo são as colunas UPOS, LEMMA (lema do token) e FEAT (“features”, ou atributos morfológicos). Cada token tem uma UPOS e, dada a UPOS, um conjunto específico de atributos morfológicos possíveis, resumidos na Tabela 1.

Tabela 1. Atributos morfológicos por UPOS

UPOS	Abbr	Case	Definite	Gender	Mood	Number	NumType	Person	Poss	PronType	Tense	VerbForm
ADJ	Yes	–	–	Fem Masc	–	Sing Plur	Ord	–	–	–	–	Part
ADP	Yes	–	–	–	–	–	–	–	–	–	–	–
ADV	Yes	–	–	–	–	–	–	–	–	–	–	–
AUX	Yes	–	–	Fem Masc	Ind Sub Imp Cnd	Sing Plur	–	1 2 3	–	–	Pres Past Fut Imp Pqp	Fin Ger Part Imp
CCONJ	Yes	–	–	–	–	–	–	–	–	–	–	–
DET	Yes	–	Def Ind	Fem Masc	–	Sing Plur	–	1 2 3	Yes	Art Ind Rel Dem Int Prs	–	–
INTJ	Yes	–	–	–	–	–	–	–	–	–	–	–
NOUN	Yes	–	–	Fem Masc	–	Sing Plur	–	–	–	–	–	–
NUM	Yes	–	–	Fem Masc	–	–	Card Frac	–	–	–	–	–
PRON	Yes	Nom Acc Dat	–	Fem Masc	–	Sing Plur	–	1 2 3	Yes	Ind Rel Dem Int Prs	–	–
PROPN	Yes	–	–	–	–	–	–	–	–	–	–	–
PUNCT	–	–	–	–	–	–	–	–	–	–	–	–
SCONJ	Yes	–	–	–	–	–	–	–	–	–	–	–
SYM	–	–	–	–	–	–	–	–	–	–	–	–
VERB	Yes	–	–	Fem Masc	Ind Sub Imp Cnd	Sing Plur	–	1 2 3	–	–	Pres Past Fut Imp Pqp	Fin Ger Part Imp
X	Yes	–	–	–	–	–	–	–	–	–	–	–

Além dos atributos descritos na Tabela 1, é possível também haver atributos não previstos em léxicos do português porque dizem respeito exclusivamente ao eixo sintagmático, como a indicação de voz passiva, **Voice=Pass**, ou porque são palavras estrangeiras, **Foreign=Yes**, ou porque integram nomes próprios, **Proper=Yes**.

3. Metodologia

Ao iniciarmos os trabalhos de anotação, encontramos tokens sem ambiguidade, tokens com ambiguidade de UPOS, de lema ou de atributos morfológicos e tokens que combinam mais de um tipo de ambiguidade. Para automatizar parte da tarefa, fizemos uso do recurso léxico PortiLexicon-UD [Lopes et al. 2022], constituído de formas da língua

portuguesa e suas respectivas possibilidades de anotação com etiquetas da UD. Os tokens que apresentavam mais de uma possível UPOS no PortiLexicon-UD foram desambiguados antes de se passar à verificação da ambiguidade de lema e de atributos morfológicos.

Por não ser nosso foco neste artigo, não detalharemos as dificuldades inerentes à resolução da ambiguidade de UPOS. Essa tarefa foi árdua, pois 44.066 tokens (26% do cópuz) apresentavam mais de uma possível UPOS. Grande parte das desambiguações de UPOS já resolveu automaticamente a anotação de lema e de atributos morfológicos, como no exemplo do token “*vestidos*” abaixo, que apresenta uma única alternativa de lema e atributos morfológicos para cada UPOS. Abaixo estão listadas as três opções de UPOS, lema e atributos morfológicos para a palavra “*vestidos*”:

- ADJ, “*vestido*”, **Gender=Masc|Number=Plur|VerbForm=Part;**
- NOUN, “*vestido*”, **Gender=Masc|Number=Plur;**
- VERB, “*vestir*”, **Gender=Masc|Number=Plur|VerbForm=Part.**

Porém, há casos em que a desambiguação de UPOS nem sempre elimina as ambiguidades de lema e de atributos morfológicos, como no exemplo do token “*fora*”, a seguir, que apresenta ambiguidades dentro das UPOS AUX e VERB.

- ADP, “*fora*”, -;
- ADV, “*fora*”, -;
- AUX, “*ser*”, **Mood=Ind|Number=Sing|Person=1|Tense=Pqp|VerbForm=Fin;**
- AUX, “*ser*”, **Mood=Ind|Number=Sing|Person=3|Tense=Pqp|VerbForm=Fin;**
- AUX, “*ir*”, **Mood=Ind|Number=Sing|Person=1|Tense=Pqp|VerbForm=Fin;**
- AUX, “*ir*”, **Mood=Ind|Number=Sing|Person=3|Tense=Pqp|VerbForm=Fin;**
- NOUN, “*fora*”, **Gender=Masc|Number=Sing;**
- VERB, “*ser*”, **Mood=Ind|Number=Sing|Person=1|Tense=Pqp|VerbForm=Fin;**
- VERB, “*ser*”, **Mood=Ind|Number=Sing|Person=3|Tense=Pqp|VerbForm=Fin;**
- VERB, “*ir*”, **Mood=Ind|Number=Sing|Person=1|Tense=Pqp|VerbForm=Fin;**
- VERB, “*ir*”, **Mood=Ind|Number=Sing|Person=3|Tense=Pqp|VerbForm=Fin.**

Sendo assim, para cada token e sua respectiva UPOS anotada no cópuz, fizemos a detecção automática de possíveis anotações de lema e atributos morfológicos segundo o PortiLexicon-UD. Anotamos automaticamente os tokens que, para uma dada UPOS, não apresentavam nem ambiguidade de lema nem de atributos morfológicos. Já para aqueles que apresentavam ambiguidade, utilizamos algumas heurísticas (descritas mais abaixo) para anotação automática e submetemos o restante à anotação manual.

É importante ressaltar que, na grande maioria dos casos, uma vez decidida a UPOS de um token, as possibilidades de atributos morfológicos já são conhecidas graças ao PortiLexicon-UD, ou seja, mesmo que haja mais de uma possibilidade de anotação, a tarefa do anotador será escolher entre as possíveis alternativas previstas no léxico.

4. Processo de Desambiguação

Percebemos, durante nossa prática de desambiguação de UPOS, que existe um efeito cascata de um tipo de desambiguação para outro, como já exemplificado acima. Como a desambiguação do lema pode resolver eventualmente as ambiguidades de atributos morfológicos, dividimos o processo em duas etapas:

- Desambiguações de lema dentro de uma mesma UPOS;
- Desambiguações de atributos morfológicos dentro de uma mesma UPOS e de um mesmo lema.

4.1. Desambiguações de Lema - mesma UPOS

Ao procurarmos tokens que admitiam mais de um lema para uma mesma UPOS, encontramos um total de 1.708 tokens, sendo 1.560 verbos plenos, auxiliares ou de cópula (VERB e AUX) que possuem **formas verbais homônimas** para verbos diferentes. As demais 148 ocorrências nessa etapa são **formas nominais homônimas** (substantivos que possuem lemas distintos).

As **formas verbais homônimas** que correspondem a conjugações de verbos distintos ocorreram em 1.560 tokens distribuídos em 1.431 sentenças. Um exemplo é a forma verbal “*viram*” encontrada nas sentenças:

- “A recepção do Neymar vocês **viram** como foi.”, que trata do verbo “*ver*”;
- “Os moradores se **viram** com grades reforçadas, câmeras e até portão novo para fechar a via.”, que trata do verbo “*virar*”.

A desambiguação dessas 1.560 ocorrências foi feita através de planilhas com as 1.431 sentenças que foram manualmente analisadas. Embora várias formas com lema verbal ambíguo tenham sido encontradas, mais da metade (841 das 1.560) eram formas comuns aos verbos “*ser*” e “*ir*”. A Tabela 2 mostra alguns exemplos de formas verbais cujos lemas foram desambiguados.

Tabela 2. Exemplos de formas verbais com lema ambíguo

Forma	Opções de Lema		Forma	Opções de Lema	
“for”	“ser”	“ir”	“fosse”	“ser”	“ir”
“dita”	“ditar”	“dizer”	“pode”	“podar”	“poder”
“postas”	“postar”	“pôr”	“traga”	“tragar”	“trazer”
“sentem”	“sentar”	“sentir”	“vira”	“virar”	“ver”

Os **substantivos com formas homônimas**, ou seja, substantivos com mais de uma possibilidade de lema, somaram 148 tokens distribuídos em 146 sentenças. Um exemplo é o token “*críticas*”, utilizado como NOUN, que pode ser o feminino plural do substantivo masculino “*crítico*” (como em “*elas são críticas de arte*”) ou pode ser plural do substantivo feminino “*crítica*” (como em “*recebeu boas críticas*”). Outro exemplo é o substantivo “*suspeita*”, que pode ser feminino singular do substantivo “*suspeito*” ou singular do substantivo feminino “*suspeita*”, como exemplificado nas sentenças a seguir:

- “Uma porta-voz da promotoria disse que a **suspeita** não fez ameaças ou declarações extremistas.”, em que o lema é “*suspeito*” ;
- “UNE processa Lollapalooza por **suspeita** de burlar lei da meia entrada.”, em que o lema é “*suspeita*”.

A desambiguação foi feita através de uma planilha com as 146 sentenças que foram analisadas manualmente. A Tabela 3 mostra alguns exemplos de substantivos cujos lemas foram desambiguados.

4.2. Desambiguações de Atributos Morfológicos - mesma UPOS e mesmo Lema

Na segunda etapa do processo, analisamos tokens que apresentam mais de uma possibilidade de atributos morfológicos, mesmo apresentando uma mesma UPOS e um mesmo lema. Foram encontrados 11.397 tokens, sendo:

- 7.543 **verbos** com UPOS VERB ou AUX;
- 3.822 **pronomes** com UPOS PRON;
- 32 **substantivos** com UPOS NOUN.

Tabela 3. Exemplos de formas nominais desambiguadas

Forma	Opções de Lema		Forma	Opções de Lema	
“ <i>crítica</i> ”	“ <i>crítico</i> ”	“ <i>crítica</i> ”	“ <i>técnica</i> ”	“ <i>técnico</i> ”	“ <i>técnica</i> ”
“ <i>química</i> ”	“ <i>químico</i> ”	“ <i>química</i> ”	“ <i>porteira</i> ”	“ <i>porteiro</i> ”	“ <i>porteira</i> ”
“ <i>mineradora</i> ”	“ <i>minerador</i> ”	“ <i>mineradora</i> ”	“ <i>música</i> ”	“ <i>músico</i> ”	“ <i>música</i> ”

4.2.1. Ambiguidades de Atributos de Verbos

Nos 7.543 tokens anotados como verbos (VERB ou AUX), a maioria (4.531 tokens) possui ambiguidade tanto nos atributos morfológicos de Pessoa (**Person**) quanto de Tempo, Modo e Forma Verbal (**Tense, Mood e VerbForm**). As ambiguidades de atributos verbais ocorrem quando diferentes conjugações de um mesmo verbo têm a mesma forma escrita (são homônimas). Por exemplo, a conjugação da segunda pessoa do singular do Imperativo é idêntica à terceira pessoa do singular do presente do Indicativo, como no verbo “*apresentar*”, cuja forma “*apresenta*” é utilizada na segunda pessoa do singular do Imperativo (“*apresenta tu*”) e na terceira pessoa do singular no presente do Indicativo (“*ele apresenta*”).

Dentro de nossa abordagem, tratamos distintamente as ocorrências onde a ambiguidade de formas verbais se dá devido a múltiplas opções de:

- Pessoa (**Person**): primeira, segunda, ou terceira;
- Número (**Number**): singular ou plural;
- Tempo, Modo e Forma Verbal, que são definidos respectivamente por três atributos em UD: **Tense, Mood e VerbForm**, os quais apresentam 14 possíveis combinações em português. A forma verbal pode ser Infinitivo (pessoal e impessoal), Gerúndio, Particípio ou Finito. Dentro das formas marcadas como Finito, temos os modos e seus respectivos tempos: Indicativo (presente, pretérito perfeito, pretérito imperfeito, pretérito mais-que-perfeito, futuro do presente e futuro do pretérito), Subjuntivo (presente, pretérito imperfeito e futuro) e Imperativo.

Desambiguação do Atributo Pessoa - Dos 7.543 verbos (VERB e AUX) com ambiguidade de atributos morfológicos, 6.379 têm ambiguidade de Pessoa (**Person**), e estes estão distribuídos em 4.580 sentenças. Um exemplo é a forma verbal “*para*” nas sentenças:

- “*A escola não **para**, as crianças estão lá todos os dias, afirmou.*”, em que “*para*” é a terceira pessoa do singular do Presente do Indicativo;
- “***Para** de chorar porque o seu marido vai cansar, relata.*”, em que “*para*” é segunda pessoa do singular do Imperativo.

Dado que o corpus é de gênero jornalístico, observamos que é mais provável que as formas ambíguas pertençam à terceira pessoa do Presente do Indicativo e não à segunda pessoa do Imperativo. Por essa razão, quando os tokens ambíguos apresentavam só essas duas opções (ou seja, em 3.782 ou 59% das ocorrências com ambiguidade de pessoa), utilizamos uma heurística para diminuir o número de casos submetidos à análise dos anotadores. A heurística consistiu em selecionar somente as sentenças onde o token ambíguo era o primeiro token da sentença, ou era precedido de aspas (“”), dois pontos (:) ou reticências (...), configurações em que o Imperativo teria maior probabilidade de ocorrer. Essa heurística resultou em 226 sentenças, contendo 318 tokens ambíguos que, analisados manualmente, revelaram 9 casos de segunda pessoa do singular do Imperativo.

Para as demais 2.789 sentenças (contendo 3.464 tokens ambíguos), atribuímos automaticamente a terceira pessoa do singular do Presente do Indicativo.

Para o restante dos verbos onde a desambiguação de Pessoa ainda era necessária, os verbos foram analisados individualmente nas 1.565 sentenças (do total de 4.580 sentenças com ambiguidade de pessoa). Um exemplo é o token “*fazia*” nas sentenças:

- “*Ela não **fazia** a menor ideia de como ou por onde começar a procurar trabalho*”, *re-corda.*”, onde “*fazia*” é a terceira pessoa do singular do Pretérito do Imperfeito do Indicativo;
- “*É uma percepção que tenho desde que **fazia** residência.*”, em que “*fazia*” é a primeira pessoa do singular do Pretérito Imperfeito do Indicativo.

Essa análise resultou na desambiguação de 2.597 tokens, dos quais 2.159 foram anotados como terceira pessoa do singular e 438 como primeira pessoa do singular. A Tabela 4 mostra alguns exemplos de tokens com ambiguidade de pessoa.

Tabela 4. Exemplos de formas verbais com ambiguidade de pessoa

Forma	Opções de Pessoa	
“ <i>demande</i> ”	primeira pessoa do sing. no Pres. do Sub.	terceira pessoa do sing. no Pres. do Sub.
“ <i>conta</i> ”	segunda pessoa do sing. do Imperativo	terceira pessoa do sing. do Pres. do Ind.
“ <i>crece</i> ”	segunda pessoa do sing. do Imperativo	terceira pessoa do sing. do Pres. do Ind.
“ <i>diz</i> ”	segunda pessoa do sing. do Imperativo	terceira pessoa do sing. do Pres. do Ind.
“ <i>absolva</i> ”	primeira pessoa do sing. no Pres. do Sub.	terceira pessoa do sing. no Pres. do Sub.

Desambiguação do Atributos Tempo e Modo - Considerando que, dos 7.543 tokens verbais com ambiguidade nos atributos morfológicos, 6.379 já foram desambiguados quando se definiu a pessoa do verbo (**Person**), restaram somente 1.013 para se desambiguar o tempo e modo do verbo (**Tense**). Esses 1.013 tokens estavam distribuídos em 714 sentenças. Um exemplo é a forma “*diga*” nas sentenças:

- “*Dá uma boa olhada em os números de a tabela e me **diga** se você não ficou com água em a boca?*”, em que “*diga*” é a terceira pessoa do singular do Imperativo;
- “*É simplesmente alguém que coloque as coisas em ordem, e **diga**: atenção, minha gente vamos nos acertar aqui e deixar as coisas de forma que o país consiga andar e não como estamos.*”, em que “*diga*” é a terceira pessoa do singular do Subjuntivo.

A desambiguação foi feita através de uma planilha com as 714 sentenças para revisão manual, resultando na anotação de 301 tokens como Presente do Subjuntivo, 207 tokens como Presente do Indicativo, 188 tokens como Pretérito Perfeito do Indicativo e 19 tokens como Futuro do Subjuntivo. A Tabela 5 mostra exemplos dessas desambiguações.

Tabela 5. Exemplos de formas verbais com ambiguidade de tempo

Forma	Opções de Tempo Verbal	
“ <i>aproveite</i> ”	terc. pess. do sing. no Imperativo	terc. pess. do sing. no Pres. do Sub.
“ <i>possam</i> ”	terc. pess. do plural do Imperativo	terc. pess. do plural do Pres. do Sub.
“ <i>tenha</i> ”	terc. pess. do sing. no Imperativo	terc. pess. do sing. no Pres. do Sub.
“ <i>precisamos</i> ”	prim. pess. do plural do Pres. do Sub.	prim. pess. do plural do Pret. do Sub.
“ <i>vão</i> ”	terc. pess. do plural do Imperativo	terc. pess. do plural do Pres. do Sub.
“ <i>mandaram</i> ”	terc. pess. do plural do Pret. Perfeito	terc. pess. do plural do Pret. Mais-que-perf.
“ <i>consequirem</i> ”	terc. pess. do plural do Fut. do Sub.	terc. pess. do plural do Infinitivo pessoal

4.2.2. Ambiguidades de Atributos de Pronomes

Um total de 3.822 tokens anotados como **pronomes** (PRON) possuem basicamente dois tipos de ambiguidade. A primeira diz respeito ao tipo do pronome (**PronType**), que apresenta ambiguidades entre os valores Relativo, Demonstrativo, Indefinido e Interrogativo, e a segunda diz respeito ao caso gramatical (**Case**): Nominativo, Dativo ou Acusativo.

Desambiguação do Atributo Tipo do Pronome - Essa ambiguidade ocorre em 2.903 tokens em 2.159 sentenças, sendo que em 2.102 deles só existe ambiguidade quanto ao tipo de pronome, enquanto os outros 801 tokens têm também ambiguidade de caso. Um exemplo de ambiguidade de tipo de pronome é o token “*quem*” nas sentenças:

- “*Até o momento, quem ganhou mais com a nova tecnologia foram os clientes.*”, em que “*quem*” é Relativo;
- “*A noite está brilhante e acetinada, quem você acha que é?*”, em que “*quem*” é Interrogativo.

Para desambiguar as 2.159 sentenças contendo os 2.303 tokens de pronome com ambiguidade de tipo, foram geradas planilhas para revisão manual. A ambiguidade de tipo de pronome ocorreu apenas sobre 11 palavras distintas que são apresentadas na Tabela 6.

Tabela 6. Exemplos de pronomes com ambiguidade de tipo

formas	Opções de Tipo de Pronomes		
“ <i>que</i> ”	Relativo	Interrogativo	–
“ <i>qual</i> ”	Relativo	Interrogativo	–
“ <i>quais</i> ”	Relativo	Interrogativo	–
“ <i>quem</i> ”	Relativo	Interrogativo	Indefinido
“ <i>quantos</i> ”	Relativo	Interrogativo	Indefinido
“ <i>tal</i> ”	Demonstrativo	Indicativo	–
“ <i>tais</i> ”	Demonstrativo	Indicativo	–
“ <i>a</i> ”	Demonstrativo	Pessoal (Caso Acusativo)	–
“ <i>as</i> ”	Demonstrativo	Pessoal (Caso Acusativo)	–
“ <i>o</i> ”	Demonstrativo	Pessoal (Caso Acusativo)	–
“ <i>os</i> ”	Demonstrativo	Pessoal (Caso Acusativo)	–

Desambiguação do Atributo Caso - Considerando que, dos 3.822 pronomes com ambiguidade iniciais, 2.903 já foram desambiguados ao decidir o tipo de pronome, restaram 919 tokens com ambiguidade de caso. Um exemplo de ambiguidade de caso é o pronome “*se*” nas sentenças:

- “*Gilmar não participou e nem o ministro Marco Aurélio, que se declarou suspeito.*”, onde “*se*” é Acusativo;
- “*As garotinhas podem se perguntar agora: eu quero ser cabeleireira ou chanceler?*”, onde “*se*” é Dativo.

A desambiguação dos 919 tokens resultou em 703 pronomes do caso Nominativo, 219 do caso Acusativo e 98 do caso Dativo. É interessante citar que essa ambiguidade ocorreu apenas para 4 pronomes distintos, todos pronomes pessoais, descritos na Tabela 7.

4.2.3. Ambiguidades de Atributos de Substantivos

O último caso de desambiguação de atributos morfológicos refere-se aos **substantivos** (NOUN), que podem ser utilizados nos gêneros Masculino, Feminino ou não apresentar

Tabela 7. Exemplos de casos de pronomes ambíguos

Palavras	Opções de Caso de Pronomes			Palavras	Opções de Caso de Pronomes	
“me”	Acusativo	Dativo	–	“te”	Acusativo	Dativo
“se”	Acusativo	Dativo	–	“nos”	Acusativo	Dativo

nenhum valor, quando são comuns de dois gêneros. Este tipo de ocorrência apareceu em apenas 32 tokens distribuídos em 32 sentenças. Por exemplo, o token “*corte*” pode ser um substantivo masculino (como em “*corte de tecido*”) ou um substantivo feminino (como em “*a decisão da corte*”), mas ambos os casos mantêm a UPOS NOUN e o lema “*corte*”. Outro exemplo de ambiguidade é o substantivo “*reservas*” nas sentenças:

- “*As chamas, que começaram nas reservas naturais, avançaram nos últimos dias com ventos de 80 km/h e chegaram às cidades.*”, em que “*reservas*” é feminino;
- “*Tite precisa definir, sem mudar o esquema tático, os reservas imediatos de Renato Augusto e Paulinho.*”, em que “*reservas*” é masculino.

As 32 sentenças foram analisadas por anotadores, resultando em 22 tokens anotados como masculino e 10 tokens anotados como feminino. Apenas 7 substantivos distintos com ambiguidade de gênero foram encontrados no corpus: “*meia*”, “*reserva*”, “*reservas*”, “*bandeirinha*”, “*corte*”, “*cortes*” e “*paquera*”.

5. Conclusão

Este estudo demonstrou a complexidade da desambiguação de lemas e atributos morfológicos em português. A metodologia empregada tem três grandes vantagens: 1) anotar automaticamente o lema e os atributos dos tokens que não apresentam ambiguidade; 2) anotar automaticamente os atributos que deixam de ser ambíguos quando uma das etapas de desambiguação é concluída; 3) sistematizar a anotação manual dos casos ambíguos, por meio de planilhas, restringindo as alternativas de anotação e diminuindo a probabilidade de erros de anotação.

A grande lição que aprendemos nesse trabalho que faz parte do Projeto POeTiSA (<https://sites.google.com/icmc.usp.br/poetisa/>) é que há um efeito de desambiguação em cascata, que justifica uma ordem de desambiguação das colunas do CoNLL-U: primeiro UPOS, depois lema e, por fim, atributos morfológicos.

O estudo evidencia o volume de tokens ambíguos que requerem uma anotação manual cuidadosa, por anotadores com bons conhecimentos linguísticos, a fim de garantir a precisão dos dados. Esperamos que este relato contribua para uma melhor compreensão dos desafios envolvidos na construção de recursos linguísticos para PLN.

Agradecimentos

Este trabalho foi realizado no âmbito do Centro de Inteligência Artificial da Universidade de São Paulo (C4AI - <http://c4ai.inova.usp.br/>), com o apoio da Fundação de Amparo à Pesquisa do Estado de São Paulo (processo FAPESP #2019/07665-4) e da IBM. Este projeto também foi apoiado pelo Ministério da Ciência, Tecnologia e Inovações, com recursos 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

- Branco, A., Silva, J. R., Gomes, L., and António Rodrigues, J. (2022). Universal grammatical dependencies for Portuguese with CINTIL data, LX processing and CLARIN support. In Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Odijk, J., and Piperidis, S., editors, *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5617–5626, Marseille, France. European Language Resources Association.
- de Marneffe, M.-C., Manning, C. D., Nivre, J., and Zeman, D. (2021). Universal Dependencies. *Computational Linguistics*, 47(2):255–308.
- Duran, M., Lopes, L., das Graças Nunes, M., and Pardo, T. (2023). The dawn of the portinari 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.
- Duran, M., Lopes, L., and Pardo, T. (2021). Descrição de numerais segundo modelo universal dependencies e sua anotação no português. In *Anais do XIII Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 344–352, Porto Alegre, RS, Brasil. SBC.
- Duran, M. S., Oliveira, H., and Scandarolli, C. (2022). Que simples que nada: a anotação da palavra que em corpus de UD. In Pardo, T. A. S., Di-Felippo, A., and Roman, N. T., editors, *Proceedings of the Universal Dependencies Brazilian Festival*, pages 1–11, Fortaleza, Brazil. Association for Computational Linguistics.
- Gamba, F. and Zeman, D. (2023). Universalising Latin Universal Dependencies: a harmonisation of Latin treebanks in UD. In Grobol, L. and Tyers, F., editors, *Proceedings of the Sixth Workshop on Universal Dependencies (UDW, GURT/SyntaxFest 2023)*, pages 7–16, Washington, D.C. Association for Computational Linguistics.
- Goldberg, Y. (2015). A primer on neural network models for natural language processing. *CoRR*, abs/1510.00726.
- Lopes, L., Duran, M., Fernandes, P., and Pardo, T. (2022). Portilexicon-ud: a portuguese lexical resource according to universal dependencies model. In *Proceedings of the Language Resources and Evaluation Conference*, pages 6635–6643, Marseille, France. European Language Resources Association.
- Lopes, L., Fernandes, P., Inacio, M. L., Duran, M. S., and Pardo, T. A. S. (2023). Disambiguation of universal dependencies part-of-speech tags of closed class words in portuguese. In Naldi, M. C. and Bianchi, R. A. C., editors, *Intelligent Systems*, pages 241–255, Cham. Springer Nature Switzerland.
- Nivre, J., de Marneffe, M.-C., Ginter, F., Goldberg, Y., Hajič, J., Manning, C. D., McDonald, R., Petrov, S., Pyysalo, S., Silveira, N., Tsarfaty, R., and Zeman, D. (2016). Universal Dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666, Portorož, Slovenia. ELRA.
- Rademaker, A., Chalub, F., Real, L., Freitas, C., Bick, E., and de Paiva, V. (2017). Universal Dependencies for Portuguese. In Montemagni, S. and Nivre, J., editors, *Pro-*

ceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017), pages 197–206, Pisa, Italy. Linköping University Electronic Press.

Souza, E., Silveira, A., Cavalcanti, T., Castro, M., and Freitas, C. (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, Porto Alegre, RS, Brasil. SBC.

Universal Dependencies (2023). CoNLL-U format - UD version 2. <https://universaldependencies.org/format.html>. Accessed: 2021-06-14.

An NLP approach to impersonal *–se* in Brazilian Portuguese

Elvis A. de Souza¹, Magali S. Duran¹, Adriana S. Pagano²

¹Interinstitutional Center for Computational Linguistics (NILC)
University of São Paulo (USP)

²Federal University of Minas Gerais (UFMG)

{elvis.desouza99, magali.duran}@gmail.com, apagano@ufmg.br

Abstract. *This paper introduces an annotation proposal for the reflexive pronoun –se in Brazilian Portuguese with a view to classifying different strategies for impersonalization through the use of one supercategory. We carried out experiments on a gold standard treebank for Portuguese in the Universal Dependencies project and verified that the implementation of our proposal results in the training of a morphosyntactic annotation model that annotates syntactic dependencies 1.27 percentage point better in accuracy. Moreover, a detailed evaluation showed an increase of up to 6.34 accuracy in the annotation of verb arguments, one of the most important classes for carrying out various Natural Language Processing tasks, highlighting the importance of informed linguistic modeling decisions in practical NLP results.*

1. Introduction

In the field of Natural Language Processing (NLP), the annotation of reflexives presents a unique challenge. Traditional grammatical approaches often fail to support analyses that can yield consistent results from a cross-linguistic perspective. This, in turn, hampers the cross-comparability that projects such as Universal Dependencies (UD) [De Marneffe et al. 2021] aim to achieve, resulting in NLP models that can not perform well in the analysis of syntactic relations. Poor performance in parsing impacts several downstream tasks that would benefit from solid annotation.

In this paper, we approach annotation of reflexives in Brazilian Portuguese through an experiment which seeks to evaluate the impact of annotating two traditionally separate uses of the pronoun *–se*, namely subject indeterminacy and synthetic passive, with a single dependency relation tag, *expl:impers*, standing for “impersonal”, given their similar characteristics. This approach is meant to offer several advantages, such as recognizing impersonalization strategies and eliminating controversial grammatical agreement issues.

Our hypothesis is that this unification will enhance automatic learning of syntactic relations implicating verb arguments, which are vital to tasks such as information extraction and Q&A. To validate our approach, we performed a conversion of tags in PetroGold [de Souza and Freitas 2023], a Brazilian Portuguese Universal Dependencies treebank with well-defined guidelines for the annotation of the pronoun *–se*, and evaluated the impact of the conversion on the quality of an NLP parsing model [Straka et al. 2016].

2. The pronoun *–se* in Brazilian Portuguese

In Brazilian Portuguese, the pronoun *–se* can have an argumental and a non-argumental role in a verb’s argument structure. It has an argumental role when it functions as direct or

indirect object in a clause. In such cases, it can construe reflexive or reciprocal meanings and is co-referential with the subject of the clause, as illustrated in examples (1) and (2)¹.

- (1) Ele não se perdoou pelo erro
 he not REFL.3 forgive-PST.PERF.3SG for the mistake
He did not forgive himself for the mistake
- (2) Eles se abraçaram depois do jogo
 they RECP.3 hug-PST.PERF.3PL after of the match
They hugged each other after the match

A non-argumental role of *-se* is implicated in two particular uses. The first involves verbs selecting a reflexive pronoun to construe a specific meaning, derived or not from a transitive meaning. This use is relevant semantically, but not syntactically. In this case, there is agreement between subject person and reflexive pronoun person, as seen in examples (3), (4) and (5).

- (3) Eu me abstive de votar
 I REFL.1SG refrain- PST.PERF.1SG from voting
I refrained from voting
- (4) O conflito se estendeu por meses
 the conflict REFL.3 extend- PST.PERF.3SG for months
The conflict dragged on for months
- (5) A plateia silenciou se
 the audience silence- PST.PERF.3SG REFL.3
The audience fell silent

A second non-argumental use of *-se* is in impersonal constructions, where it can only operate as a third person pronoun for verbs in third person singular or plural forms. When the verb selecting *-se* is intransitive (6) or indirect transitive (7), *-se* is considered a marker of subject indeterminacy and the verb has a third person singular form. However, when the verb is direct transitive, *-se* is considered a marker for passive voice and its complement to the right classified as a passive subject, which requires it to agree in number with the verb, as seen in examples (8) and (9).

3. Reflexives in the Universal Dependencies framework

Within the UD framework, reflexives are annotated following their distinction in terms of whether they have an argumental role or not. Argumental reflexives are pronouns realizing objects (tagged *obj* or *iobj*) and in some languages obliques (*obl*). Non-argumental

¹Examples include a gloss and a free translation to assist readers, particularly when *-se* does not translate into a corresponding pronoun in English.

- (6) Caiu se numa armadilha
 fall-PST.PERF.3SG REFL.3 into a trap
We fell into a trap
- (7) Pensou se em tudo
 think-PST.PERF.3SG REFL.3 in everything
We thought of all available possibilities
- (8) Utilizou se uma única teoria
 use-PST.PERF.3SG REFL.3 a single theory
We used a single theory / A single theory was used
- (9) Levantaram se várias hipóteses
 raise- PST.PERF.3PL REFL.3 several hypotheses
We raised several hypotheses / Several hypotheses were raised

reflexives are categorized as expletives. The latter include three types of expletives for reflexive markers: (1) markers of impersonal meanings (tagged *expl:impers*), (2) reflexive pronouns construing passive meanings (tagged *expl:pass*) and (3) reflexive pronouns selected by pronominal verbs (tagged *expl:pv*).

Given their extensive use and frequency in several language families, reflexives have been target of several studies, both as one among other expletive categories and as a category in itself. All studies point out inconsistencies in annotation with impact on UD's much desired cross-comparability among treebanks. [Bouma et al. 2018], for instance, discuss reflexives within the broad class of expletives and advocate current UD guidelines. [Marković and Zeman 2018] focus on reflexive markers and perform a cross-linguistic comparison of patterns in three language families: germanic, romance and slavic. They explore UD treebanks and find many inconsistencies in their annotation. They strongly advocate annotating reflexive markers by selecting “Yes” for the Reflex feature as a form of quickly querying treebanks for these markers. Regarding reflexive expletives, they advocate prioritizing the annotation of inherent reflexives (*expl:pv*), on the one hand, and impersonal and passive reflexives (*expl:impers* and *expl:pass*), on the other.

[Degraeuwe and Goethals 2020] also report problems in treebank annotation consistency, focusing on Spanish in particular. They advocate using existing UD tags for argumental (*obj*, *iobj*) and non-argumental (*expl:pv*, *expl:impers*, *expl:pass*) roles, but propose annotating fine-grained distinctions regarding reflexive and non-reflexive functions by means of feature properties.

4. A proposal for simplifying non-argumental and non-reflexive *-se* in Portuguese

Annotating the three subrelations proposed by UD guidelines in Brazilian Portuguese texts is actually a very hard and time-consuming task, as it involves lengthy discussions which sometimes do not achieve a consensus among annotators. This can seriously impact consistency in treebanks, a problem reported by the above-mentioned studies. There are

also considerations which challenge some of the assumptions underlying the expletive tags.

The tag *expl:pass* implicates a passive voice reading, which is not the case in Brazilian Portuguese, as the verb form is not passive. Moreover, these constructions do not admit expression of a passive agent, as is the case in passive voice constructions. A further issue in Brazilian Portuguese is increasing loose agreement between the verb form and the complement to the right, which raises discussions as to whether the complement is an object or a passive subject, as illustrated by example (10).

- (10) Vê se muitas crianças sozinhas na rua
 see-PRS.3SG REFL.3 many children alone in the street
You can see many children alone on the streets

In fact, [Magalhães and Carvalho 2021] analyzed a collection of sentences from contemporary Brazilian newspapers and found that variation in subject-verb agreement for synthetic passives is around 50%, both with finite verbs and in the infinitive.² According to traditional grammar [Bechara 2012, Bechara 2018, Cunha and Cintra 2016], the lack of agreement would be considered a grammatical mistake, since in passive constructions there must be subject-verb agreement. We conclude that the phenomenon of synthetic passives is in the process of merging with that of indeterminacy in Brazilian Portuguese, a thesis also supported by authors such as [Bagno 2012], justifying the choice to merge them in the annotation as well.

Considering these cases as strategies for construing impersonal meanings would allow for strengthening comparability with other languages which have active voice constructions for similar meanings, as is the case of pronouns “you” and “one” in English, “on” in French and “uno” in Spanish.

The above considerations led us to propose to use a single tag, namely *expl:impers*, to annotate cases formerly annotated as *expl:pass*. Several advantages ensue from this proposal. Impersonal active voice constructions of the kind we have illustrated will be treated as impersonalization strategies and their complement considered an object. Agreement issues between verb form and complement will be solved as both third person singular and plural forms will be equally annotated as objects. Constructions with a direct transitive verb selecting *-se* and having a clausal complement (*cf.* example 11) will be considered impersonal constructions as well. Finally, passive voice (tagged Voice=Pass) will be a feature assigned exclusively to past participle constructions, both in reduced and fully expanded clauses.

- (11) Via se que todos estavam desesperados
 see- PST.PERF.3SG REFL.3 that everyone was desperate
You could see that everyone was in despair

²Some examples from the analysis include: “A urgência de se aprovar leis imprescindíveis” (lit. “The urgency of *passing*(singular) essential laws(plural)”) and “não se pode perder os amigos” (lit. “one can’t lose(singular) their friends(plural)”).

5. Methodology

In order to demonstrate the advantages of our proposal for the purposes of treebank annotation, we carried out experiments using PetroGold, a gold standard treebank comprising academic text in the oil and gas domain [de Souza et al. 2021] (8,946 sentences, 250,605 tokens). We chose PetroGold as currently this is the only dataset in the UD project (v2.14) that clearly distinguishes the different uses of the pronoun *—se*.³

PetroGold annotates non-argumental uses of the pronoun *—se* with UD labels *expl:impers*, *expl:pass* and *expl:pv*. Annotation draws on traditional grammars of Brazilian Portuguese, which classify the pronoun *—se* as (i) a marker of subject indeterminacy (annotated as *expl:impers*), (ii) a pronoun for passive constructions (annotated as *expl:pass*) and (iii) a pronoun selected by pronominal verbs (*expl:pv*). These three cases are illustrated by examples (12), (13) and (14) retrieved from PetroGold.

(12) *expl:impers*: Quando **se fala** em PHPA, **fala se** em a união de a acrilamida e de o acrilato de sódio através de a copolimerização.⁴

(13) *expl:pass*: Para a síntese **usou se** manta aquecida, temperatura de refluxo e agitação mecânica.⁵

(14) *expl:pv*: Este estudo **se baseia** em as propriedades magnéticas de os minerais que **se concentram** em as rochas de a crosta terrestre.⁶

In order to join occurrences of *—se* labeled as *expl:pass* and *expl:impers* in PetroGold, we took the following steps:

- 1) The *expl:impers* label was retained in all annotated cases.
- 2) Cases previously annotated with *expl:pass* were annotated as *expl:impers* (Figure 1 exemplifies an impersonal meaning previously annotated as *expl:pass*)

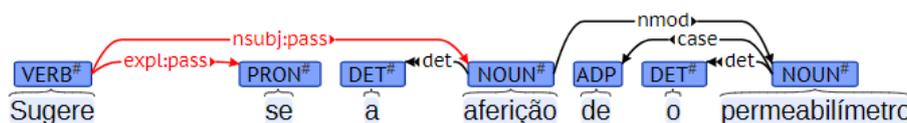


Figure 1. Original annotation of a sentence using *expl:pass*

- 3) The feature *Voice=Pass* was reserved for passive voice constructions, with a past participle form and passive voice auxiliary “ser”.
- 4) Subjects of impersonal constructions previously annotated as *expl:pass* (*nsubj:pass*) were annotated as objects of active voice constructions (*obj*) (Figure 2 shows changes made to example in Figure 1). The *nsubj:pass* tag was hence reserved for passive voice constructions, with a past participle form.

³Bosque [Rademaker et al. 2017] contains a residual annotation of *expl:pass* for only 4 verb lemmas selecting the pronoun; PUD [Zeman et al. 2017] annotates all *—se* pronouns as *expl:pv*, CINTIL [Branco et al. 2022] makes no use of any UD *expl* tags to annotate the pronoun *—se*, and Portinari [Duran et al. 2023] annotates all non-argumental *—se* as *expl*.

⁴Transl. “When we **refer** to PHPA, we **are referring** to the union of acrylamide and sodium acrylate through copolymerization.”

⁵Transl. “For the synthesis, a heated blanket, reflux temperature and mechanical stirring **were used**.”

⁶Transl. “This study **is based** on the magnetic properties of the minerals that **are concentrated** in the rocks of the Earth’s crust.”

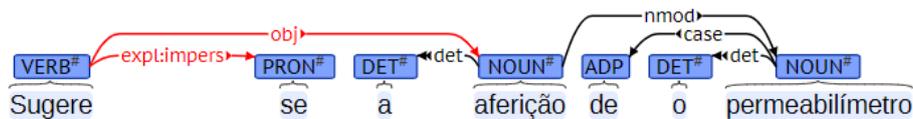


Figure 2. Current annotation of a former *expl:pass* sentence as *expl:impers*

To evaluate the quality of an NLP parsing model trained on the dataset after implementing our conversion, we trained two annotation models based on UDPipe 1.2.0 [Straka et al. 2016], one before and the other after the conversion, without any change in the model hyperparameters. The models were evaluated using the CLAS (content-label attachment score) metric, which indicates the correctness of syntactic relations and dependencies attachment for words with lexical content [Zeman et al. 2018], and the differences are shown as absolute percentage points.

Furthermore, to detail each of the tags impacted by this modification to the tagset, we used the metrics UAS (unlabeled-attachment score) and LAS (labeled-attachment score), which measure the hit rate of syntactic dependencies attachment ignoring the relation label and considering it, respectively [Zeman et al. 2018]. Other metrics are UPOS (universal part-of-speech), UFeats (universal morphological features) and Lemmas (correct prediction of lemmas), which are not directly related to the experiments conducted but may present minor impact due to the nature of neural networks training in a pipeline fashion [Straka et al. 2016].

6. Results

The treebank originally contained a total of 578 subjects (nsubj:pass) of impersonal constructions previously labeled as *expl:pass*, which were converted into active voice objects of verbs, and a total of 793 pronouns –*se* annotated with the label *expl:pass*, which were re-annotated as *expl:impers*. In total, 1,371 tokens were impacted, which corresponds to 0.5% of the tokens in the treebank. In terms of number of sentences in the corpus, 710 sentences were re-annotated, which corresponds to 7.9% of the sentences in the corpus.

Table 1 shows the results for the evaluation metrics of the trained model after the conversions were performed. Δ is the difference between the current performance value and the value before the conversions were performed. We can see a slight improvement in all metrics related to morphological features (UPOS, UFeats, Lemmas). Regarding these metrics, the only expected difference would be “UFeats”, since we removed the feature *Voice=Pass* from finite verb forms previously categorized with it. Metrics related to syntactic analysis (UAS, LAS and CLAS) show larger improvement, of up to 1.27 CLAS increase – a considerable increase, taking into account that only 0.5% of the tokens were impacted in the corpus.

To examine the parser’s performance for each of the linguistic categories in the dataset, we inspected the annotation model’s predictions on the PetroGold test partition (1,039 of the 8,946 sentences, or 11.6%). This partition contains 147 labels of type *expl:impers* resulting from the conversion of *expl:pass* and 78 objects (obj) resulting from the conversion of passive subjects (nsubj:pass).

Table 2 indicates the performance of the trained parser for each of the treebank’s syntactic categories, where “Freq.” stands for the number of times the label appears in

Metric	%	Δ
UPOS	98.43	0.02
UFeats	98.17	0.09
Lemmas	98.81	0.04
UAS	91.04	0.60
LAS	89.38	0.77
CLAS	84.57	1.27

Table 1. Evaluation metrics for the model trained on the new annotation

the test portion of the dataset, “Hits” for the number of times the parser annotated the dependency relation correctly, and “LAS” when it predicted the correct dependency attachment as well. The three last columns indicate the difference when comparing the data and the model performance before and after the conversions, and column Δ Freq. confirms that only tokens with relations *nsubj:pass*, *obj* and *expl:impers* were re-annotated. For space reasons, we display in the table only categories that have increased or decreased performance in a number higher than 1.27, which is the increase in CLAS of the model.

Relation	Freq.	Hits	LAS	Δ Freq.	Δ Hits	Δ LAS
acl	592	93.58%	79.90%	0	0.84	2.37
acl:relcl	212	88.21%	68.87%	0	-2.83	0.47
advmod	746	93.43%	79.49%	0	-0.81	1.61
appos	234	79.06%	61.11%	0	-1.71	0.85
aux	40	72.50%	70.00%	0	-10.00	-10.00
cc	765	97.52%	88.50%	0	0.53	1.83
ccomp	140	75.00%	74.29%	0	2.86	3.58
conj	1025	88.98%	60.78%	0	0.39	3.02
cop	339	96.76%	92.63%	0	2.07	1.77
expl:impers	224	89.73%	89.29%	147	68.95	68.51
expl:pvt	71	92.96%	91.55%	0	4.23	4.23
flat:name	181	83.98%	83.43%	0	-1.10	-1.65
nsubj:pass	380	88.68%	88.16%	-78	0.69	1.48
nummod	496	94.96%	92.34%	0	1.41	1.82
obj	819	93.04%	91.33%	78	5.46	5.50
obl	1486	85.94%	78.60%	0	1.55	1.88
obl:agent	102	95.10%	95.10%	0	1.96	1.96
obl:arg	142	49.30%	48.59%	0	6.34	5.63
parataxis	119	77.31%	68.91%	0	0.00	3.36
xcomp	298	87.58%	87.58%	0	1.67	2.01

Table 2. Performance of the annotation model for some dependency tags

To verify our initial hypothesis, i.e., that merging impersonal types of *-se* is promising for improving results for NLP tasks such as question and answer (Q&A) and information extraction, results for specific categories need to be taken into account. These were verb arguments – *nsubj* (+0.57), *nsubj:pass* (+1.48), *obj* (+5.50), *obl:arg* (objects introduced by preposition, +6.34), *ccomp* (objects in form of subordinate clauses, +3.58). Hence, annotation by the trained model after merging impersonal *-se* showed improve-

ment for all classes that implicate arguments, even those that were not directly affected in the gold annotation.

One of the classes that showed the highest improvement was *obj*, which can be accounted for by the fact that the proposed annotation labels arguments to the right of verbs associated with the pronoun *-se* as objects, with no more room for the model to consider them subjects. Another class with a considerable increase was *ccomp*, which can also be accounted for by the fact that constructions with a direct transitive verb selecting *-se* and having a clausal complement are not passive constructions, as formerly considered, and their complement should be tagged as *ccomp*.

The class *obl:arg*, used for verbal arguments introduced by prepositions, increased 6.34 points. Although an improvement in verbal arguments was expected, prepositional arguments were not impacted in the dataset, since only direct objects (former patient subjects) were targeted.⁷ Moreover, a class which had a 10 point decrease in the model performance in comparison to the model trained before our experiment was *aux*, used in Brazilian Portuguese to construe tense through auxiliary verbs (“*ter*”, “*haver*”, “*ir*”, “*estar*”). Such large differences merit further investigation, being possibly due to the parser neural network training procedure, which happens in a pipeline fashion, yielding no explainability for results.

7. Conclusion

Constructions with *-se* traditionally seen as construing passive meanings in Brazilian Portuguese pose a challenge to annotators, as they behave differently from regular passive constructions. They do not allow for the expression of an agent; they take a complement to the right of the verb, a position typically filled by objects; they have an active verb form, and exhibit an increasingly loose agreement between verb and complement. Our annotation considers these constructions as impersonalization strategies which can be readily annotated by means of the *expl:impers* tag. Our results show improvements in the model’s recognition of classes and our proposal is expected to allow for quicker decision-making and better consistency in treebank annotation.

There are some limitations, though, to the proposal implemented and evaluated in our study. One is the fact that there are no occurrences of pronoun *-se* as an object (*obj*) or an indirect object (*iobj*) in PetroGold, which restricts the scope of the analysis. Another one is the fact that no extrinsic evaluation was performed to verify the impact of the proposed simplification. Future work should be done to evaluate the effect of the proposed annotation of the pronoun *-se* in NLP applications.

Acknowledgments

This work was carried out at the Center for Artificial Intelligence, University of São Paulo (<http://c4ai.inova.usp.br/>), with support of São Paulo Research Foundation (FAPESP grant 2019/07665-4) and IBM Corporation. The project was supported by the Ministry of Science, Technology and Innovation with resources of Law 8248, October 23, 1991, within the scope of PPI-SOFTEX, coordinated by Softex and published as Residence in TIC 13, DOU 01245.010222/2022-44. ASP has a grant from CNPq (313103/2021-6).

⁷Prepositional objects are tagged as *obl:arg* in UD, which roughly correspond to indirect objects in traditional grammars.

References

- [Bagno 2012] Bagno, M. (2012). *Gramática pedagógica do português brasileiro*. Parábola Ed.
- [Bechara 2012] Bechara, E. (2012). *Moderna gramática portuguesa*. Nova Fronteira.
- [Bechara 2018] Bechara, E. (2018). *Lições de português pela análise sintática*. Nova Fronteira.
- [Bouma et al. 2018] Bouma, G., Hajic, J., Haug, D., Nivre, J., Solberg, P. E., and Øvrelid, L. (2018). Expletives in Universal Dependency treebanks. In de Marneffe, M.-C., Lynn, T., and Schuster, S., editors, *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, pages 18–26, Brussels, Belgium. Association for Computational Linguistics.
- [Branco et al. 2022] Branco, A., Silva, J., Gomes, L., and Rodrigues, J. (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.
- [Cunha and Cintra 2016] Cunha, C. and Cintra, L. (2016). *Nova gramática do português contemporâneo*. LEXIKON Editora Digital Ltda.
- [De Marneffe et al. 2021] De Marneffe, M.-C., Manning, C. D., Nivre, J., and Zeman, D. (2021). Universal dependencies. *Computational linguistics*, 47(2):255–308.
- [de Souza and Freitas 2023] de Souza, E. and Freitas, C. (2023). Explorando variações no tagset e na anotação universal dependencies (ud) para português: Possibilidades e resultados com base no treebank petrogold. In *Anais do XIV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 125–134. SBC.
- [de Souza et al. 2021] de Souza, E., Silveira, A., Cavalcanti, T., Castro, M., and Freitas, C. (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, Porto Alegre, RS, Brasil. SBC.
- [Degraeuwe and Goethals 2020] Degraeuwe, J. and Goethals, P. (2020). Reflexive pronouns in Spanish Universal Dependencies. *PROCESAMIENTO DEL LENGUAJE NATURAL*, 64(64):77–84.
- [Duran et al. 2023] Duran, M. S., Lopes, L., Nunes, M. d. G. V., and Pardo, T. A. S. (2023). The dawn of the porttinari multigenre treebank: introducing its journalistic portion. *Anais*.
- [Magalhães and Carvalho 2021] Magalhães, H. L. P. and Carvalho, H. M. d. (2021). Uso variável da concordância verbal em construções de voz passiva sintética na escrita de textos jornalísticos cearenses.
- [Marković and Zeman 2018] Marković, S. and Zeman, D. (2018). Reflexives in universal dependencies.
- [Rademaker et al. 2017] Rademaker, A., Chalub, F., Real, L., Freitas, C., Bick, E., and De Paiva, V. (2017). Universal dependencies for Portuguese. In *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017)*, pages 197–206.

- [Straka et al. 2016] Straka, M., Hajic, J., and Straková, J. (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.
- [Zeman et al. 2018] Zeman, D., Hajic, J., Popel, M., Potthast, M., Straka, M., Ginter, F., Nivre, J., and Petrov, S. (2018). Conll 2018 shared task: Multilingual parsing from raw text to universal dependencies. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual parsing from raw text to universal dependencies*, pages 1–21.
- [Zeman et al. 2017] Zeman, D., Popel, M., Straka, M., Hajic, J., Nivre, J., Ginter, F., Luotolahti, J., Pyysalo, S., Petrov, S., Potthast, M., Tyers, F., Badmaeva, E., Gokirmak, M., Nedoluzhko, A., Cinkova, S., Hajic jr., J., Hlavacova, J., Kettnerová, V., Uresova, Z., Kanerva, J., Ojala, S., Missilä, A., Manning, C. D., Schuster, S., Reddy, S., Taji, D., Habash, N., Leung, H., de Marneffe, M.-C., Sanguinetti, M., Simi, M., Kanayama, H., dePaiva, V., Drozanova, K., Martínez Alonso, H., Çöltekin, c., Sulubacak, U., Uszkoreit, H., Macketanz, V., Burchardt, A., Harris, K., Marheinecke, K., Rehm, G., Kayadelen, T., Attia, M., Elkahky, A., Yu, Z., Pitler, E., Lertpradit, S., Mandl, M., Kirchner, J., Alcalde, H. F., Strnadová, J., Banerjee, E., Manurung, R., Stella, A., Shimada, A., Kwak, S., Mendonca, G., Lando, T., Nitisaroj, R., and Li, J. (2017). Conll 2017 shared task: Multilingual parsing from raw text to universal dependencies. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–19, Vancouver, Canada. Association for Computational Linguistics.

Inferências baseadas em sintaxe: a anotação de sujeitos implícitos

Magali Sanches Duran, Maria das Graças Volpe Nunes, Thiago Pardo

ICMC - Universidade de São Paulo, São Carlos, SP
Núcleo Interinstitucional de Linguística Computacional - NILC
Center for Artificial Intelligence - C4AI

magali.duran@gmail.com, gracan@icmc.usp.br, taspardo@icmc.usp.br

Abstract. *This paper presents rules for annotating implicit subjects based on five syntactic dependency relations from the Universal Dependencies (UD) approach. Two of these rules were proposed by UD and improved for Portuguese, and the other three were originally developed for Portuguese and can be applied to other languages with minor adaptations.*

Resumo. *Este artigo apresenta regras para a anotação de sujeitos implícitos a partir de cinco relações de dependência sintática da abordagem Universal Dependencies (UD). Duas dessas regras foram propostas pela UD e aperfeiçoadas para o português, e as outras três foram desenvolvidas originalmente para o português, podendo ser aplicadas a outras línguas com pequenas adaptações.*

1. Introdução

Identificar o sujeito de um predicado talvez seja uma das principais motivações para uma aplicação utilizar um *parser* em seu pré-processamento, sendo importante, por exemplo, para aplicações de extração de informação, sumarização e perguntas e respostas. Com exceção dos verbos impessoais (ex: haver, chover), todo verbo tem potencialmente um sujeito. No entanto, atendendo ao princípio pragmático da economia, os mecanismos das línguas permitem suprimir sujeitos que possam ser inferidos pelos seus falantes. Por exemplo, em “João estava tentando acalmar Luíza e convencê-la a não desistir de viajar”, sabemos que “João” é sujeito de “tentar”, “acalmar” e “convencer” e Luíza é sujeito de “desistir” e “viajar”, mas apenas um sujeito está anotado sintaticamente: “João”, como sujeito de “tentar”. Em suma, temos 5 verbos e 1 sujeito (não contamos “estava”, que é verbo auxiliar).

Mecanismos de economia de sujeitos explícitos ocorrem em todas as línguas, mas são ainda mais frequentes em línguas *pro-drop*, como o português, que admitem a elipse do sujeito. Comparando “João disse que chega amanhã e que pretende jantar conosco” com sua tradução para o inglês “João said he will arrive tomorrow and that he wants to have dinner with us”, observamos que as duas sentenças têm quatro verbos, mas, enquanto o português tem apenas um sujeito explícito, o inglês tem três.

Como o sujeito elíptico de uma oração quase sempre pode ser identificado em outra oração, dentro da mesma sentença, nos últimos anos desenvolveu-se um tipo de anotação que utiliza regras para, a partir da pré-anotação sintática, inferir e anotar sujeitos não explícitos. Essa tarefa está compreendida dentro do escopo conhecido como *enhanced dependencies* (ED). As ED foram desenvolvidas primeiro no inglês (Schuster

& Manning, 2016), depois foram generalizadas para outras línguas pela abordagem *Universal Dependencies* (UD)¹ (Nivre et al., 2018, de Marneffe et al. 2021) e instanciadas para o português (Pagano et al., 2023).

Na UD foram estabelecidas regras para atribuir sujeitos de orações coordenadas e de orações subordinadas objetivas diretas e indiretas reduzidas, as quais envolvem as relações de dependência conj e xcomp. Assumindo que toda oração tem potencialmente um sujeito, analisamos todos os demais tipos de orações da UD para avaliar a oportunidade de construir novas regras de ED para atribuição de sujeitos (acl, acl:relcl, advcl, ccomp, csubj). As orações csubj (sujeito oracional) e acl:relcl (orações relativas) não apresentam potencial para isso, mas acl (orações adjetivas), advcl (orações adverbiais) e ccomp (orações objetivas diretas e indiretas desenvolvidas) apresentam.

Discutiremos o refinamento das regras das ED universais para inferência de sujeitos (nas relações UD conj e xcomp) e apresentaremos regras para novas ED concebidas (nas relações UD ccomp, acl e advcl). Para obter os resultados aqui discutidos, analisamos sentenças do *corpúsculo Portinari-base* (Duran et al. 2023) que não tinham um sujeito próprio anotado (*corpúsculo disponível no site do Projeto POeTiSA*²). Em cada caso, respondemos a uma pergunta: o sujeito da oração pode ser inferido dentro da própria sentença? Os “sim” nos deram regras para inferências corretas, e os “não” nos deram regras para evitar inferências incorretas. A Seção 2 deste artigo apresenta os tipos de ED; a Seção 3 mostra as relações da UD que podem ter sujeitos implícitos e a Seção 4 traz as conclusões do estudo.

2. *Enhanced Dependencies*

Enhanced dependencies (ED), no projeto UD, são relações inferidas a partir das relações dependências sintáticas básicas. Para exemplificar, tomemos a sentença “*João acordou e saiu*”. Nas dependências básicas (em preto), “João” é nsubj de “acordou” (Fig. 1a). Já nas ED (em vermelho), “João” é nsubj de “acordou” e de “saiu” (Fig. 1b).

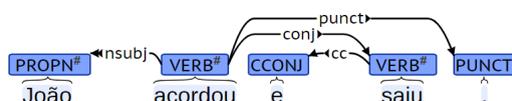


Figura 1a. Dependências básicas da sentença “João acordou e saiu”

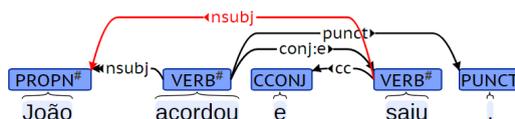


Figura 1b. *Enhanced dependencies* da sentença “João acordou e saiu”

A diferença entre as duas anotações é que, nas ED, foi acrescentada a relação nsubj entre “saiu” e “João”, relação não permitida nas dependências básicas porque cada dependente deve ter um único head. Além disso, produz um cruzamento de arcos indesejado, pois dificulta o aprendizado automático.

¹ São chamadas *Universal Enhanced Dependencies* (EUD). Como nossa proposta inclui, mas extrapola, as EUD, usamos simplesmente *Enhanced Dependencies* (ED).

² <https://sites.google.com/icmc.usp.br/poetisa>

3. Explicitação de sujeito com *enhanced dependencies*

Apresentamos a seguir cinco relações de dependência (xcomp, conj, ccomp, advcl e acl), a partir das quais se pode revelar sujeitos não anotados sintaticamente. As regras que envolvem xcomp e conj foram propostas pela UD e detalhadas neste trabalho para contemplar com maior precisão todos os casos de língua portuguesa. As regras para as outras três são contribuições deste trabalho. Usamos o termo “atribuição” para sujeitos inferidos cuja inserção no texto não é gramaticalmente aceitável (xcomp e acl) e o termo “propagação” para sujeitos inferidos cuja inserção não fere a gramaticalidade.

3.1. Atribuição de sujeito de xcomp

A relação xcomp tem como característica o fato de que seu dependente não admite um sujeito explícito. Essa ausência de sujeito é chamada *null subject* ou “sujeito nulo”. Apesar de o dependente de xcomp não ter um sujeito sintático, é possível inferi-lo, pois o sujeito nulo é “controlado” por um *token* presente na oração *head* de xcomp. A seguir são ilustrados diferentes *tokens* assumindo o controle do sujeito da xcomp: o sujeito (nsubj) (Fig. 2); o objeto direto (obj) (Fig. 3) ou, em casos mais raros, o objeto indireto (iobj, obl) (Fig. 4 e 5). As ED são destacadas em vermelho.

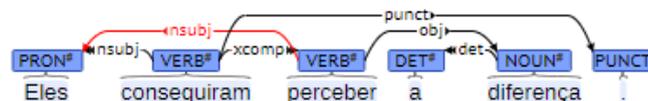


Figura 2. Sujeito da xcomp = sujeito: “Eles conseguiram perceber a diferença”

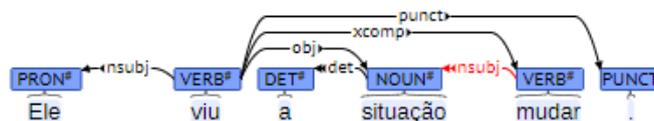


Figura 3. Sujeito da xcomp = obj: “Ele viu a situação mudar”

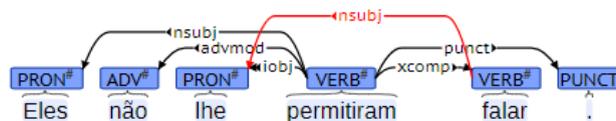


Figura 4. Sujeito da xcomp = iobj: “Eles não lhe permitiram falar”

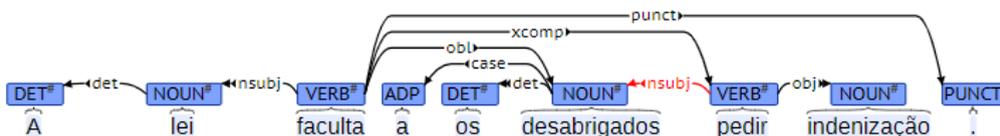


Figura 5. Sujeito da xcomp = obl: “A lei faculta aos desabrigados pedir indenização”

Se o *head* da xcomp está na voz ativa e o dependente da xcomp está na voz passiva, é necessário adequar o nome da relação de dependência, de nsubj para nsubj:pass (ou vice-versa), a fim de refletir isso, como mostra a Figura 6.

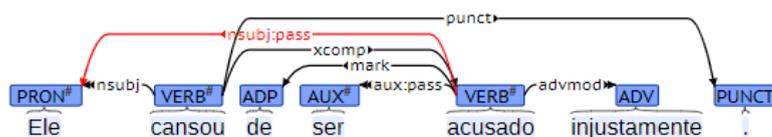


Figura 6. Inversão para voz passiva: “Ele cansou de ser acusado injustamente”

Mesmo quando o dependente de xcomp é um predicado nominal, a ED opera normalmente, como pode ser observado na Fig. 7.

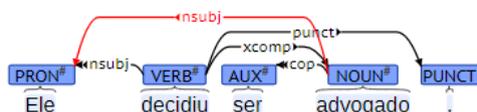


Figura 7. Dependente de xcomp predicado nominal: “Ele decidiu ser advogado”

Se, no entanto, o *token* que deveria controlar o sujeito de xcomp não estiver explícito, o dependente de xcomp não terá um sujeito atribuído, como nos exemplos a seguir, nos quais o *token* controlador elíptico ou indeterminado está indicado por um \emptyset . Embora na primeira sentença o sujeito esteja marcado na pessoa do verbo (“eu” comecei), não existe um *token* que possa receber a relação de sujeito.

- [\emptyset] Comecei a trabalhar cedinho.
- [\emptyset] Pretende-se chegar até amanhã.
- O chefe mandou [\emptyset] fazer plantão.

Verbos que preveem o objeto da oração matriz como controlador do sujeito do xcomp pertencem a classes restritas (causativos, resultativo, de percepção, com predicativo do objeto), como nos quatro exemplos a seguir.

- Ele **nos** mandou/deixou/fez esperar. (“nos”=nós, sujeito de “esperar”)
- Ele teve a **casa** invadida. (“casa”, sujeito da passiva de “invadida”)
- Ele viu/ouviu/sentiu a **terra** tremer. (“terra”, sujeito de “tremer”)
- Ele acha/considera/julga **isso** impossível. (“isso”, sujeito de “impossível”)

3.2 Propagação de sujeito do *head* para dependente de conj

Se dois predicados estão ligados por conj, o sujeito do *head* pode ser propagado para o dependente, observadas as condições (vide Seção 3.6 para condições gerais para propagação de sujeito de conj, ccomp e advcl). A Figura 8 ilustra essa propriedade para o dependente do tipo nsbj na sentença “Essa **música** é **alegre** e **animada**”.

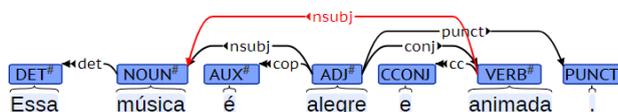


Figura 8. Propagação da relação nsbj para o dependente de conj

As Figuras 9, 10 e 11 ilustram casos de propagação de sujeito da ativa (nsbj), sujeito da passiva (nsbj:pass) e sujeito oracional (csbj), respectivamente.

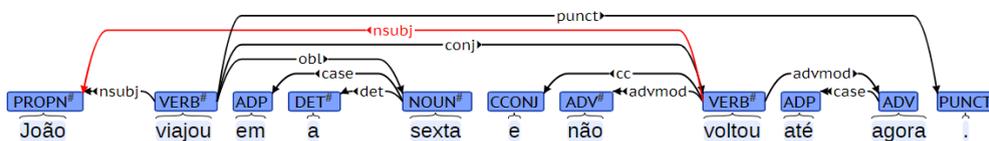


Figura 9. Propagação de nsbj em “João viajou na sexta e não voltou até agora”

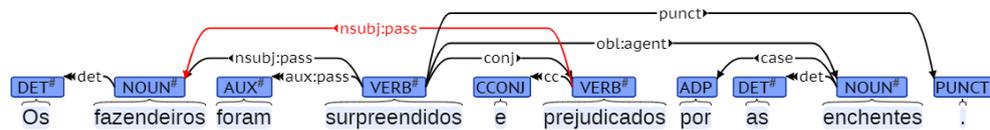


Figura 10. Propagação de nsubj:pass em "Os fazendeiros foram surpreendidos e prejudicados pelas enchentes"

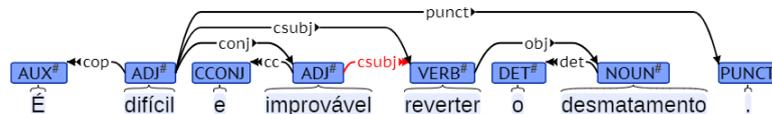


Figura 11. Propagação de csubj em "É difícil e improvável reverter o desmatamento"

Se o *head* de conj não tiver sujeito ("Acordei cedo e logo saí para caminhar") ou se o dependente de conj já tiver seu próprio sujeito ("A comida era gratuita e a bebida era barata"), não haverá sujeito a ser propagado.

Quando a relação conj liga duas orações, pode haver inversão de voz ativa para passiva ou vice-versa, conforme ilustra a Fig.12.

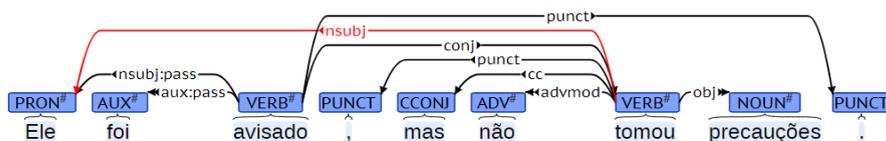


Figura 12. Propagação do sujeito em "Ele foi avisado, mas não tomou precauções"

Também é possível "aproveitar" o sujeito atribuído numa ED para fazer outra ED. Isso ocorre na Figura 13 ("Ele pretende estudar engenharia e trabalhar na construção de navios"), onde duas orações coordenadas, sem sujeito ("estudar" e "trabalhar"), ocorrem depois de uma oração xcomp que teve seu sujeito (Ele) atribuído em uma outra ED.

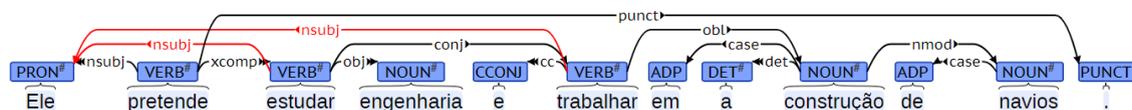


Figura 13. Propagação de sujeito de xcomp para head e dependente de conj

3.3. Propagação de sujeito do *head* para dependente de ccomp

No português, as orações subordinadas do tipo ccomp podem ter o sujeito elíptico e esse sujeito pode coincidir com o sujeito da oração *head* de ccomp, por isso é possível criar uma ED propagando o sujeito do *head* de ccomp para o dependente de ccomp. Vide regras de propagação gerais na Seção 3.6.

Essa propagação de sujeito de ccomp colocaria as línguas *pro drop* em condições de igualdade com as línguas que não admitem a elipse de sujeito. As Figuras 14a e 14b ilustram a propagação de um sujeito de ccomp no português ("Ele disse que [ele] vai aposentar") e equivalente em inglês (*He said he will retire*), língua na qual o sujeito do dependente de ccomp não pode sofrer elipse e, portanto, não precisa da ED.

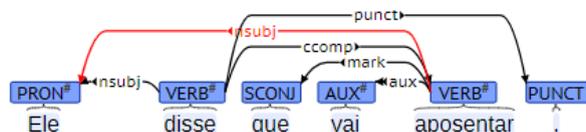


Figura 14a. Propagação de sujeito de ccomp no português

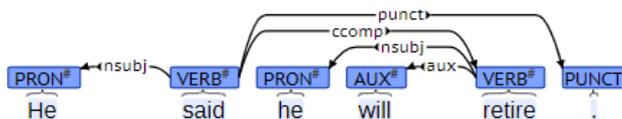


Figura 14b. Sujeito de ccomp explícito no inglês

A propagação do sujeito para preencher o sujeito elíptico do dependente de ccomp tem outra vantagem: não interrompe a anotação de cadeias de sujeitos implícitos. Na sentença "Ele disse que [ele] vai aposentar e [ele] pretende [ele] viajar durante um ano" (Figura 15), a anotação do sujeito do dependente de ccomp ("aposentar") torna possível propagar o mesmo sujeito para a oração dependente de conj ("pretende") e anotar o sujeito controlador da oração dependente de xcomp ("viajar").

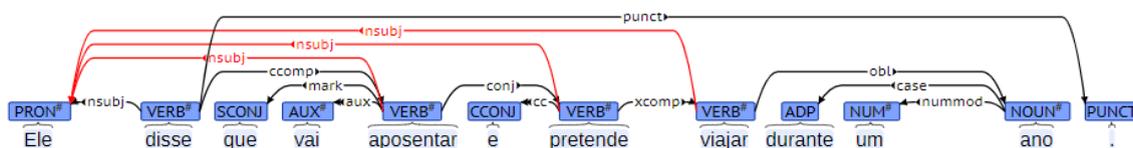


Figura 15. Sentença com ccomp com propagação do sujeito

Se um dos predicados envolvidos na relação ccomp tiver Voice=Pass e o outro não, adapta-se a relação de nsubj:pass para nsubj ou nsubj para nsubj:pass.

3.4 Propagação de sujeito do *head* para dependente de advcl

As orações subordinadas do tipo advcl também podem apresentar elipse de sujeito e, portanto, são candidatas a obter a propagação do sujeito de seus respectivos *heads* (vide regras de propagação gerais na Seção 3.6). Contudo, como há vários tipos de orações adverbiais, muitas delas com várias elipses, as regras para propagação do sujeito nem sempre alcançam uma boa precisão. Nos exemplos a seguir, o sujeito do *head* não propaga para o dependente da advcl, embora as regras básicas tenham sido atendidas.

Quando **acordou**, o veículo não estava mais **lá**. (“veículo” é sujeito do predicado nominal “lá”, mas não é sujeito de “acordou”)

Minha intenção é **contribuir** com isso, **trazendo** dados confiáveis. (“intenção” é sujeito de “contribuir”, mas não é sujeito de “trazendo”)

Não temos estatísticas para avaliar o percentual de casos que fogem à regra (isso será feito quando as regras forem implementadas computacionalmente e os casos revisados). As Figuras 16 e 17 ilustram respectivamente uma advcl temporal e uma advcl concessiva com propagação do sujeito.

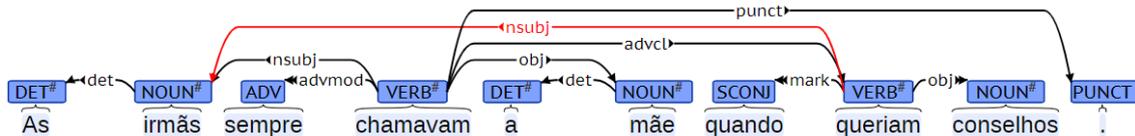


Figura 16. Propagação de sujeito de advcl "As irmãs sempre chamavam a mãe quando queriam conselhos."

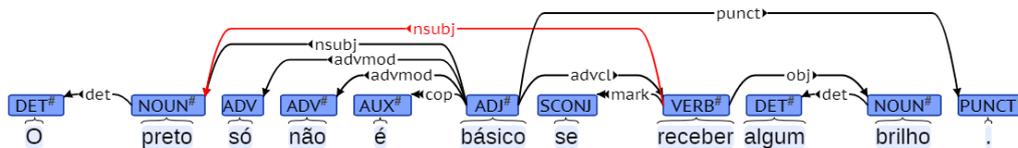


Figura 17. Propagação sujeito de advcl "O preto só não é básico se receber algum brilho."

3.5 Atribuição de sujeito de acl

Os dependentes de acl são orações adjetivas ou completivas nominais e apresentam forma de orações reduzidas, isto é, estão no particípio, gerúndio ou infinitivo. A relação acl une um *token* substantivo a uma oração. Em algumas condições, o *token head* da acl é também o sujeito lógico do dependente da acl e por isso poderia ser atribuído nas ED.

As acl cujo dependente é um **verbo no particípio** são orações adjetivas reduzidas de voz passiva. Por isso, o *token head* da relação acl será nsubj:pass do dependente, como ilustra a Figura 18 na sentença "A **certeza** de vitória **demonstrada** pelos advogados de defesa é surpreendente" (= "a certeza que foi demonstrada").

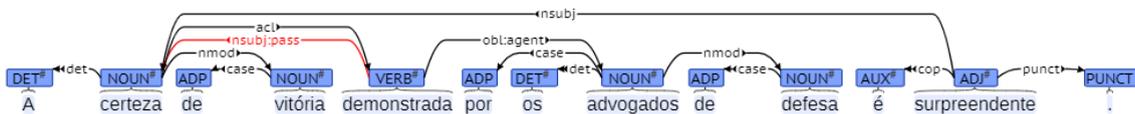


Figura 18. Anotação de nsubj:pass de dependente de acl no particípio

As acl **reduzidas de gerúndio** são sempre adjetivas e podem, seguramente, receber atribuição de sujeito, como na sentença da Figura 19 "Recebi uma **mensagem** **dizendo** que a carga tombou." (= "mensagem que dizia"). Nesse caso, o sujeito é nsubj.

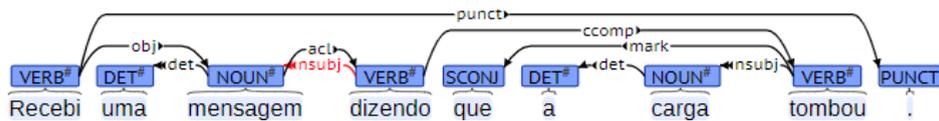


Figura 19. Anotação de nsubj de dependente de acl no gerúndio

As acl **reduzidas de infinitivo** podem constituir orações adjetivas ("máquina de lavar roupa", "preço a combinar") ou orações completivas nominais ("vontade de chorar", "interesse em viajar"). Quando são adjetivas, o *head* da acl pode ser anotado como nsubj ou como nsubj:pass do dependente da acl, como mostram, respectivamente, o primeiro e o segundo exemplo a seguir.

Ela foi a primeira **mulher** a **assumir** o cargo. (= mulher que assumiu o cargo)

Esse é o **preço** a **pagar** por um jantar ali. (= "preço que será pago")

As acl do tipo complemento nominal não obedecem essa regra (o *head* da acl não é nsubj ou nsubj:pass do dependente). Algumas dessas orações não admitem atribuição de sujeito, pois denotam uma indeterminação do sujeito ("Propina é a **maneira** mais fácil de [se] **atravessar** um muro na fronteira"); outras têm, como sujeito lógico, o sujeito da oração em que se encontra o *head* da acl ("Eles têm **vontade** de viajar").

3.6 Regras gerais para a propagação de sujeitos de conj, ccomp e advcl

As regras para propagar sujeito do *head* para o dependente de conj, ccomp e advcl são:

- o *head* de conj, ccomp ou advcl precisa ter um sujeito explícito (nsubj, nsubj:pass, csubj, csubj:pass);
- o predicado dependente de conj, ccomp ou advcl não deve ter sujeito próprio;
- o predicado *head* e o predicado dependente de conj, ccomp ou advcl têm que ter a mesma pessoa e número (condição estendida para o auxiliar ou verbo de cópula caso o predicado os tenha como dependentes);
- o predicado dependente de conj, ccomp ou advcl não pode ser um verbo impessoal (por exemplo: "haver", "chover");
- o predicado dependente de conj, ccomp ou advcl não pode estar impessoalizado pelo índice de indeterminação do sujeito "se" (expl:impers, na anotação UD).

4. Conclusões

Discutimos o potencial de multiplicação de anotação de sujeitos sintáticos inferíveis por meio de regras construídas a partir das relações de dependência sintática UD, relações chamadas de *enhanced dependencies*. Todos os tipos de orações da UD foram avaliados e, portanto, não enxergamos possibilidade de novas EDs de sujeito nesse momento. Algumas dessas regras foram definidas no âmbito do projeto UD e aperfeiçoadas para o português neste estudo (sujeitos de xcomp e conj). Outras foram levantadas a partir da observação de contextos do português, mas são igualmente aplicáveis a outras línguas, incluindo o inglês (sujeitos de advcl e acl), ou exclusivamente a línguas que admitem elipse do sujeito, como espanhol e italiano (sujeito de ccomp). As regras descritas serão utilizadas em um programa para automatização da anotação de *enhanced dependencies* no português, seguindo a abordagem simbólica já adotada por outras línguas, conforme relatos de duas *shared tasks* dedicadas ao assunto [Bouma et al., 2020; 2021]. Esse programa será utilizado para anotar o *corp*us Portinari-base (Duran et al. 2023).

Agradecimentos

Este trabalho foi realizado no âmbito do Centro de Inteligência Artificial da Universidade de São Paulo (C4AI -<http://c4ai.inova.usp.br/>), com o apoio da Fundação de Amparo à Pesquisa do Estado de São Paulo (processo FAPESP #2019/07665-4) e da IBM. Este projeto também foi apoiado pelo Ministério da Ciência, Tecnologia e Inovações, com recursos 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

- Bouma, G., Seddah, D., and Zeman, D. (2020). Overview of the iwpt 2020 shared task on parsing into enhanced universal dependencies. In 58th Annual Meeting of the Association for Computational Linguistics.
- Bouma, G., Seddah, D., and Zeman, D. (2021). From raw text to enhanced universal dependencies: The parsing shared task at iwpt 2021. In Proceedings of the 17th International Conference on Parsing Technologies and the IWPT 2021 Shared Task on Parsing into Enhanced Universal Dependencies (IWPT 2021), pages 146–157. Association for Computational Linguistics (ACL).
- Duran, M. S. (2024). Anotação de Enhanced Dependencies: Orientações para anotação de relações de dependência sintática do tipo enhanced em Língua Portuguesa, seguindo as diretrizes da abordagem Universal Dependencies (UD). Relatório Técnico do Núcleo Interinstitucional de Linguística Computacional (NILC). Disponível em: <https://repositorio.usp.br/item/003209188>
- Duran, M. S.; Lopes, L.; Nunes, M.G.V.; Pardo, T.A.S. (2023). The Dawn of the Portinari Multigenre Treebank: Introducing its Journalistic Portion. In the Proceedings of the 14th Symposium in Information and Human Language Technology (STIL), pp. 115-124. September, 25-29. Disponível em: <https://sol.sbc.org.br/index.php/stil/article/view/25443/25264>
- de Marneffe, M.-C.; Manning, C.D.; Nivre, J.; Zeman, D. (2021) Universal Dependencies. *Computational Linguistics* 47(2), 255-308.
- Nivre, J.; Marongiu, P.; Ginter, F.; Kanerva, J.; Montemagni, S.; Schuster, S.; Simi, M. (2018) Enhancing Universal Dependency Treebanks: A Case Study. In Proceedings of the Second Workshop on Universal Dependencies, pages 102-107.
- Pagano, A. S.; Duran, M. S.; Pardo, T. A. S. (2023) Enhanced dependencies para o português brasileiro. In: Proceedings of the 2nd Edition of the Universal Dependencies Brazilian Festival, p. 461–470, Belo Horizonte, Brasil. Association for Computational Linguistics. Disponível em: <https://sol.sbc.org.br/index.php/stil/article/view/25486/25307>
- Schuster, S.; Manning, C. D. (2016). Enhanced English Universal Dependencies: An Improved Representation for Natural Language Understanding Tasks. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), pages 2371–2378, Portorož, Slovenia. European Language Resources Association (ELRA).

Notes on variation and lexical diachrony in the *Parish Memories-Alentejo* collection (1758)

Helena Freire Cameron¹, Fernanda Olival², Renata Vieira²

¹CIDEHUS- Portalegre Polytechnic University, Portugal, Portugal

²CIDEHUS-University of Évora, Portugal

helenac@ipportalegre.pt, mfo@uevora.pt, renatav@uevora.pt

Abstract. *Memórias Paroquiais-Alentejo (1758)* collects the responses of the parish priests from the largest region of Portugal (Alentejo) to a survey carried out by the Crown, asking about the state of the territory and its populations, and also about the effects of the earthquake 1755. This article discusses the transformative process from the manuscripts up to the processable digital stage. We described some individualising characteristics of the lexical “corpus” obtained with PLN. Marks of lexical diachrony also stand out, revealing society and its uses in the mid-18th century.

Resumo. *As Memórias Paroquiais-Alentejo (1758)* reúnem as respostas dos párocos da maior região de Portugal (Alentejo) a um inquérito feito pela Coroa, sobre o estado do território e das suas populações, e também sobre os efeitos do sismo de 1755. Neste artigo dá-se conta do processo transformativo dos manuscritos até à fase digital processável. Descrevem-se algumas características individualizadoras do corpus lexical realizadas através de PLN. Destacam-se, igualmente, marcas de diacronia lexical, reveladoras da sociedade de então e dos seus usos.

1. *Parish Memories: constitution and support transfers*

In order to obtain a better knowledge of the territory, the Portuguese crown, in 1758, through the Secretariat of State for Business of the Kingdom, carried out a survey of the entire continental territory of Portugal with 60 questions. He asked about the “land”, the “mountain”, the “river”, and the effects of the Lisbon earthquake of 1755. He tried to collect information about the territory, the buildings, the populations and aspects that today we would classify as material culture, such as religious festivities and local memory, based on identity factors, notable aspects and illustrious characters. This initiative had precedents. There was previous information resulting from other surveys, namely the one in 1732 and other subsequent ones (Capela (1998)). However, the earthquake not only destroyed much of this information, such as that collected by the Oratorian Father Luís Cardoso using a geographical dictionary, but also changed the country’s reality (Cosme & Varandas (2009)). Therefore, the crown launched a new survey, almost certainly with the help of the invoked Father. It should reach every parish in the kingdom. Each formed the smallest administrative unit in the territory, the parish, which then had only a religious reference.

The parish priests' responses were later gathered into a documentary collection, aiming to continue the creation of a future geographical dictionary initiated by Father Luís Cardoso (Bigotte Chorão (1987)). The collection was bound in the 19th century and is now in the custody of the National Archives of Torre do Tombo (ANTT) in Lisbon. It would later become known as *Memórias Paroquiais*, a name both “consecrated and incorrect”, according to Bigotte Chorão (2023), since it deals with the parish priests' responses to the inquiry and not the memories of the clergy themselves, as the title may suggest.

Given the high demand for consulting of this valuable collection, between 1993 and 2003, ANTT microfilmed the originals and digitized them in 2005, making them available online, in image format. Therefore, the face-to-face consultations of these species have ended on paper.

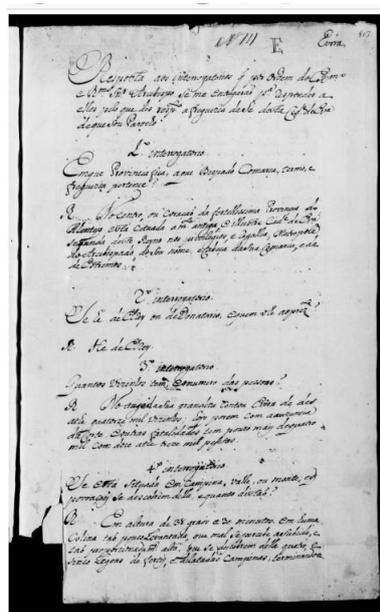


Figure 1: Extract from the *Parish Memory* of Évora, p. 1, ANTT, PT-TT-MPRQ-14-111, at <https://digitalq.arquivos.pt/viewer?id=4240009>

In 2007, the CIDEHUS research centre (Interdisciplinary Center for History, Cultures and Societies of the University of Évora) began transcribing the *Parish Memoirs* of Alentejo (southern area of Portugal), onto digital media in a fully processable format. Three years later, part of the collection was made available online and later migrated to the data aggregator repository CIDEHUSDigital (www.cidehusdigital.uevora.pt), where the texts can be searched, almost always in 18th-century spelling (with some interventions by transcribers/proofreaders). A search engine is also available, enabling more subtle searches (Santos et al. (2020)).

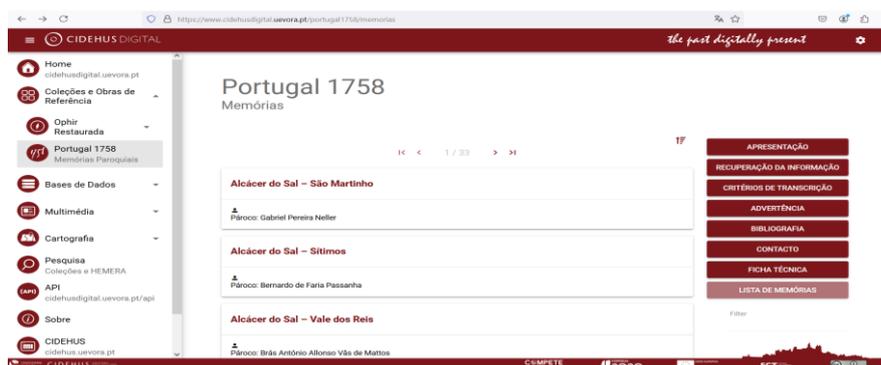


Figure 2: Collection *Portugal 1758*, at [CIDEHUSDigital](https://www.cidehusdigital.uevora.pt/pt/pt/1758/memorias)

2. Brief notes from the lexical *corpus* of *Parish Memories-Alentejo*

In 2020-2022, a multidisciplinary team from CIDEHUS-University of Évora created a lexical corpus with the 366 Memórias do Alentejo, relating to the parishes that belong to the current districts of Portalegre, Évora and Beja. The texts were obtained from the transcriptions available on CIDEHUSDigital. When transcribing, fidelity to the original was preserved whenever possible, maintaining the spelling and punctuation of the handwritten originals. Even so, there were some interventions, namely: some transcribers updated “u” with a consonant value, registering it as “v”, abbreviations were developed, some capital letters that were inserted in the middle of the word were updated, and the long “s”.

This *corpus* consists of texts of various dimensions, written by priests of various ages with various backgrounds and geographic and social origins. These factors make this *corpus* representative of the Portuguese language in this region, in the mid-1700s, and unique in the linguistic panorama of the 18th century in Portugal.

The *corpus* from the 366 Memories mentioned, non-lemmatized and non-standardized, contains 34,181 simple words and 632,498 occurrences, processed using the *AntConc* tool (version 4.0.3). An alphabetical list and a list ordered by descending frequency were prepared. In this article, we indicate the number of occurrences for each word cited in parentheses.

Excluding stop words, some of the most frequently used words are “villa” (4621), “freguezia” (2737), “senhora” (2476), “igreja” (2254), “senhor” (1683), “legoas” (1558), “cidade” (1469), “anno” (1342), “santo” (1208), “rio” (1206), “serra” (1150), “dista” (1141), “altar” (994). These words reveal the purpose of the surveys and reflect the parish priests' vision of the region, with a high incidence of words linked to religious experience.

Concerning linguistic characteristics, graphic variation is a characteristic of this period of the Portuguese language, already abundantly described in the literature (Roberts and Kato (1993), Kemmler (2001), Gonçalves (2003), Castro (2006), Cardeira and Mateus (2008), Martins (1995, 2016, ...), Banza and Gonçalves (2018), Galves, Brito and Paixão de Sousa (2005), Galves, Kato and Roberts (2019), among others equally relevant). The *corpus* of Memórias Paroquiais-Alentejo highlights, in an expressive way, the graphic variation, not only as a characteristic of the 18th-century period, but also due to the plural constitution of the *corpus* itself, Cameron (2023).

Linguistic variation in the *corpus* has several types (Cameron, 2023). Here, we present some examples of this variation, which is not always predictable and often incoherent. At the level of double consonants without linguistic criteria, forms with and without a doubled consonant coexist, with different numbers of occurrences of each of the variants. In the *corpus*, we have “abbade”(19), but there are no occurrences without the double consonant. Nevertheless, the priests wrote “abbobada”(1), “abobada”(45), “abobeda”(220) and, in plural, “abobadas”(5) and “abobedas”(25). In this last example, the variant with the highest number of occurrences does not correspond to the current word, singular or plural.

Often, the same word contains registers of variation of different phenomena, with variants in terms of doubled consonants and the register of the nasal diphthong, for example in “cappelam”(6), “cappellam”(7), “cappellao”(1), “cappelão”(10), “cappellão”(26), “cappellám”(1), “cappellã”(2), “cappelã”(1), and, in plural, “cappellaens”(1), “cappellaenz”(1), “cappellaes”(3), “cappellães”(1), “cappeloens”(1).

There are many variants with a single occurrence, almost half the number of simple words in the *corpus*. These singularly occurring variants often demonstrate lapses or errors on the part of the writers or even some lesser knowledge of the language. We note some of them: “addemiravens”(1) (for admiráveis), “apouzentado”(1) (for aposentado), “Bultiau”(1) (referring the lexicographer Bluteau), “coditiana”(1) (for cotidiana), “domecilio”(1) (for domicílio), “espisial”(1) (for especial), “espriencias”(1) (for experiências), “exziste”(1) (for existe), “hipicondriacos”(1) (for hipocondríacos), “obestujiaõ”(1) (for obstrução), among many others that are equally interesting.

The *corpus* also contains numerous Latin words, whether expressions of religious use, such as “sede vacante”, “authoritate ordinarii”, or even Latin expressions, such as “ostiatim”, “in perpetuum”, among others (See, in this regard, about the *Memories* of Vila Viçosa, Olival (2023)).

We morphologically annotated the *corpus* using the *Lx Tagger tool*, available in [Portulan Clarin](#). Although the *corpus* was not previously normalised, the tool recognised and correctly labelled most of the words. We used the labels that were pre-defined by the tool.

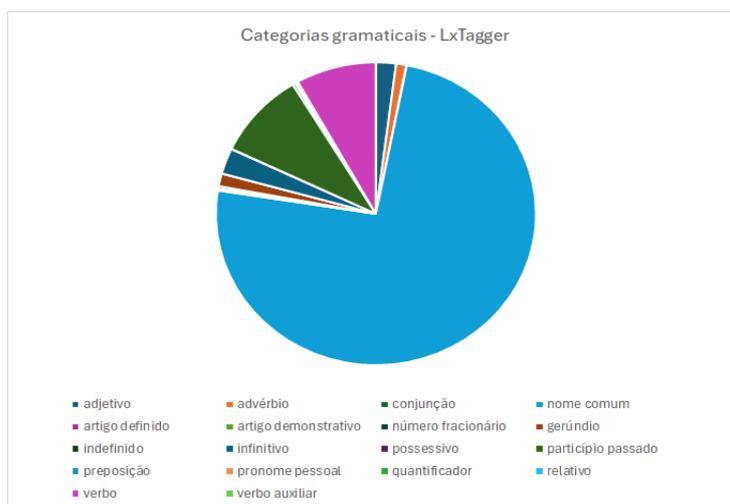


Figure 3: Breakdown of grammatical categories in the non-lemmatized and non-standardized *corpus*

Most of the *corpus* comprises nouns, followed by past participles and, thirdly, verbs. It should be noted that the *corpus* was not previously lemmatised, so the results obtained must be seen from this perspective.

3. Some lexical diachrony marks

The *corpus* of *Memórias Paroquiais-Alentejo*, in addition to being an “x-ray” of the Portuguese language from the 18th century, also constitutes a privileged historical guide to the lexicon relating to Portuguese society, especially in Alentejo, its occupations, uses and habits. Some lexical expressions reveal this past experience and are unknown to current speakers, as they have unusual meanings in the 21st century. Thus, in a second phase of the *corpus* study, the researchers read the texts, highlighting terms that had historically limited meanings or no longer used today. Once the terms were highlighted, they were observed in the *AntConc* tool, using the KWIC (keywords in context) function, allowing the word to be seen in context and giving immediate access to the entire original text.

Likewise, lexicographic existences of some of the terms were searched in previous or contemporary dictionaries: in the dictionaries of Jerónimo Cardoso (16th century), Bento Pereira (7th edition, from the end of the 17th century), Rafael Bluteau (from the first quarter of the 18th century). We also consulted António de Morais Silva's dictionary, in its 10th edition, from the mid-20th century, in 12 volumes. In 21st century dictionaries, we used the Houaiss dictionary, in its first Brazilian edition, 2001.

The *corpus* is rich in very interesting examples. We chose three of them, which clarify this diachrony, revealing characteristics of society in this region at that time:

3.1. “Herdade de cavalaria” [TRAD: Cavalry homestead]

The expression “cavalry estate” is curious. The meaning used has nothing to do with the etymological root, horse, but refers to an economic occupation of the territory beginning to appear in this region at that time. In the entire *corpus*, it appears in only three *Memories*, which we reproduce:

“[...] mas como nesta freguesia andão muitas **herdades de cavalaria**, ha muytos cazeiros, e como cada cazeiro, /p. 1183/ está em costume de pagar ao Prior hum alqueire de trigo, sempre o Prior fas certos quatro mojos de trigo”.

Monforte-Nossa Senhora da Graça, ANTT, *Memórias Paroquiais*, vol. 24, nº 179, p.1182-1183

The second Memory clarifies in detail the meaning of the term:

“[...] 27º Constando esta freguezia de cincoenta herdades, somente tem dezouto lavradores; e as outras trinta e duas herdades andão **de cavalaria**, e destas se achão treze herdades com as cazas, ou moradias derribadas, a fim de que algum lavrador não pretenda orar nellas; algumas forão destruidas no tempo da guerra, e outras tem sido derribadas pelos mesmos, que as cultivão **de cavalaria**, em grave damno, e prejuizo da Republica, e do bem commum; porquanto os lavradores, que trazem as **herdades da cavalaria**, não as cultivão, nem lavrão, como erão cultivadas, e lavradas naquelle tempo, em que em cada huma herdade morava seo lavrador, e lavrava as suas folhas enteyras; o que agora não fazem, nem cultivão, ou lavrão, a fim de terem pastajes largas para criar ovelhas, e venderem muntas láas aos estrangeiros, que as levão para o Norte.

Vila Viçosa - Nossa Senhora das Ciladas, ANTT, *Memórias Paroquiais*, vol. 11, nº 326, p. 2234.

In the third Memory, the parish priest warns of the dangers of this type of economic occupation:

“[...] e podendo recolher muytos mais mantimentos os não recolhem por se utelizarem das terras, para pastajem dos gados, por cujo interesse tem deyxado cahir a mayor parte dos montes, que há nesta freguezia, que tendo legoa, e meya de comprimento, e outro tanto de largura; e trinta e sinco herdades, se achão só /p. 434/ habitando nella, sette lavradores, e as mais andão **de cavalaria** por cujo motivo falta o culto Devino, mantimentos no Reyno, sustento aos Parochos;”

Elvas - Senhora da Ajuda, ANTT, *Memórias paroquiais*, vol. 1, nº 62, pp. 433-434

The lexicographic register of this expression is not systematic. Bluteau, contemporary of the parish priests who wrote the *Memories*, already notes this meaning:

“cavallaria [...] Andar de Cavallaria. No Alentejo quando arrendão as herdades, huma das clausulas he, que não andem de cavallaria, querem dizer, que sejaõ habitadas pelos arrendadores, porque sem habitadores se perdem, e cahem as casas.”

Bluteau, R. (1712-1728), *Vocabulario*, vol. 9.

In the 20th century, Morais (1949-1959) registers the expression in more detail and it is closer to the meaning used in the *corpus* of the Memories.

“[...] *herdade de cavalaria*, herdade que não tem cultura ou se cultivava só uma pequena parte || Pôr, trazer a herdade de cavalaria, cultivar a menor parte e deixar o resto a monte [...]”

Morais Silva, A. (1949-1959) *Grande Dicionário da Língua Portuguesa*, vol II, p. 1043.

Houaiss (2001) marks the expression as being a Lusitanianism, only used in Portugal:

“cavalaria [...] P. Herdade que não tem cultura ou que cultivava apenas uma parte de suas terras.

Houaiss (2001), *Dicionário Houaiss da Língua Portuguesa*, p. 659.

The meaning was not found in other contemporary dictionaries.

3.2. “Polícia” [TRAD: Police]

The term “polícia” has two occurrences in the *corpus* under consideration, both in meanings very distant from contemporary common usage. The two occurrences are included in the same Memory, which means the same person wrote both forms. However, there seem to be some differences between them:

“[...] Fora das grades do Cruzeiro sta a capella de Sancto Antonio, que com o quadro do mesmo Sancto mandou fazer o Excelentíssimo Dom Balthasar de Faria de marmores finissimos de varias cores primorisamente lavrados, onde se acha o nome de seu Excelentíssimo author gravado em huma pedra em letras inniciaes douradas enlaçadas com muita arte, e **policia**;”

Elvas-Sé, ANTT, *Memórias Paroquiais*, vol. 13, nº (E) 14, p. 76

“[...] Correndo os annos se lhe ajuntarão companheiros com mais **policia** e noticia de letras, que se ordenarão em presbíteros.”

Elvas-Sé, ANTT, *Memórias Paroquiais*, vol. 13, nº (E) 14, p. 71 a 106, p. 98

Morais (1949-1959) includes two meanings that seem to coincide with the uses given by the parish priest of the Cathedral of Elvas:

[...] apuro, esmero, cuidado ou correção da linguagem [...] || Delicadeza, aprumo, correção, polidez, civilidade [...] || Conjunto de preceitos técnicos destinados a um dado fim [...]

Morais Silva, A. (1949-1959), *Grande Dicionário da Língua Portuguesa*, vol. VIII, p. 444

The first occurrence seems equivalent to a set of technical precepts while the second may be equivalent to “care”.

We did not find, in contemporary dictionaries, meanings coinciding with the 18th century use found in the *Parish Memories*. However, this use of the term “policia” has a lexicographic record prior to the 18th century, in main bilingual dictionaries: Jerónimo Cardoso, Bento Pereira, and Bluteau:

“policia. policies(ei). politura(ae).”

Cardoso, Jerónimo (1562), *Dictionarium ex lusitanico in latinum*.

“Policia. Cicilitas, atis,. Urbanitas, atis. Elegantia, ae.”

Pereira, Bento (1647) *Thesouro da Língua Portuguesa*.

“[...] Policia no trato, na conversaçãõ, nos costumes, etc. Urbanitas, ou comitas, atis. Fem. Cic. Morum elegantia, ae. Fem. Tacit. {(A honra he a fonte de todo o bom ensino, Policia, procedimento, etc. Lobo, Corte na Aldea, Dialog. 15. pag. 306.)} {(A brandura, & Policia de vossas palavras. Idem. Primavera, 3. part. 194.)} § Policia, tambem se toma pela boa graça nas ações, & gestos do corpo, etc. Vid. Garbo, Graça, etc. {(Vereis alguns que fallaõ às pancadas, & se acharem hü pulpito diante, o farãõ em pedaços, como se a Policia podera soffrer o desassossego, & inquietaçãõ da sua esgrima. Lobo, Corte na Aldea, Dialog. 8. pag. 172.)} § Policia, algumas vezes val o mesmo, que Aceyo, Limpeza, Alinho, etc. Vid. nos seus lugares. {(A brandura no conversar, a Policia no vestir, a cortezania no tratar, Lobo, Dialog. 5. mihi pag. 114.)} {(Lhe servissem com grande Policia ministros escolhidos as melhores iguarias. [...])”

Bluteau, R. (1712-1728) *Vocabulario*, vol.6.

3.3. “País” [TRAD: Country]

The word country is used in the *Memories* in two different meanings, designating a territory with a political organization, a use that is predominant today:

“[...] indo sahindo da igreja o tomarão os mouros, e captivo o levãrão para Argel, sem mais pessoa alguma deste **país**.”

Vila Nova de Milfontes - Nossa Senhora da Graça, ANTT, *Memórias Paroquiais*, vol. 23, nº 142, p. 908

and referring to a specific region, as is common today in the French language:

“[...] He o seu curso neste **País** moderado, sem mi[ni]ma precipitação, por serem planicie, os campos por onde passa.”

Elvas-Terena, ANTT, *Memórias Paroquiais*, vol. 36, nº 44, p. 277

“[...] laranja doce deste **pays** he particular no gosto em todo Alentejo [...]”

Montemor-o-novo-Santiago do Escoural, ANTT, *Memórias Paroquiais*, vol. 14, nº 55, p. 402

Houaiss (2001) includes this meaning, referring it to the French term “pays”, which has been used since the Renaissance. This meaning can be recognized in derived words such as “paisagem” (landscape), “paisagismo” (landscaping), “paisano”, among others.

Bluteau, in the middle of a definition for another purpose, uses the term in this sense, referring to “region”:

“Aquitania. Grande provincia de França, que os modernos divide em tres, a saber a primeyra Aquitania, que contem o Berry, o pays de Bourbon, a alta, & a baxa Alvernia,

o pays de Velay, de Gevedan, de Roverga, dos Albigenses, de Querey, de Limoges, & da alta, & baxa Macha. A 2. Aquitania contem o pays de Bordeos [...]
Bluteau, R. (1712-1728), *Vocabulario*, vol.I.

This meaning had already been included in the 7th edition of *Prosodia* (1697), by Bento Pereira:

“Adrana, ae, m. g. Pais de Alemanha. p. l. ex Graec. Tacit. lib. 1. ann. [...]”
Pereira, Bento, (1697) *Prosodia*, 7th ed.

4. Conclusions

The lexical *corpus* of Memórias Paroquiais-Alentejo brings together an expressive graphic variation, characteristic of the time, and also the result of the multiplicity of authors who wrote the almost four hundred texts that constitute it. It contains lexical diachrony marks, as we saw briefly. It is a true source from a linguistic point of view. Access to the lexical mass, with the support of concordance generation tools, allows each word to be seen in context, enabling the construction of lexical datasets, which can feed ontologies and, at the same time, constitute observatories of societal changes (e.g. “cavalry estate”).

Given the size of the *corpus*, this first phase of defragmenting the text into simple words, useful for studying variation, was followed by a second phase of morphological annotation of the *corpus*. Almost two-thirds of the *corpus* is filled with names. However, it is estimated that this value may decrease after lemmatization.

The third task was carried out manually, observing the texts, in order to see the words in context. The *corpus* is being annotated with Named Entities customised to the historical context (Vieira, 2021), (Santos, 2024). However, the study of these low-frequency terms cannot be carried out automatically, always requiring the human intervention of linguists and historians. Therefore, the authors did a new systematic reading of the texts, to withdraw uses and meanings with diachronic value.

The automation of historical *corpora* will greatly benefit from all the efforts to process and study the Portuguese lexicon. However, a subsequent phase of human intervention will always be necessary, to fully enhance the study and dissemination of these valuable historical lexical *corpora*. The power of the Artificial Intelligence arm is still limited.

5. References

- Banza, A. P. et Gonçalves, M. F. (2018) *Roteiro de História da Língua Portuguesa*. Évora: Cátedra UNESCO - Universidade de Évora.
- Bigotte Chorão, M. J. M. (2023) «As Memórias Paroquiais de 1758, de novo», *Notícias da Torre do Tombo*. Lisboa, N.º 1 (jul. 2023).
- Bigotte Chorão, M. J. M. (1987) “Inquéritos promovidos pela coroa no século XVIII”, *Revista de História Económica e Social*, nº 21, Lisboa, pp. 93-130.
- Bluteau, R. (1712-1728) *Vocabulario Portuguez e Latino*, Coimbra, Lisboa, Lisboa Colégio das Artes, Pascoal da Sylva, Joseph Antonio da Sylva, Patriarcal Officina da Musica.

- Cameron, H.F, Olival, F., Vieira, R. (2023) Planear a normalização automática: tipologia de variação gráfica do corpus das Memórias Paroquiais (1758), *LaborHistorico*, v.9 n.1, <https://doi.org/10.24206/lh.v9i1.52234>
- Capela, J. V. (1998) *Barcelos nas Memórias Paroquiais de 1758*, Câmara Municipal de Barcelos.
- Cardeira, E. et Mateus, M. H. M. (2008) *Norma e Variação*. Alfragide: Editorial Caminho.
- Cardoso, J. (1569-1570) *Dictionarium latinolusitanicum & vice versa lusitanicolatinum*, Coimbra, João de Barreira.
- Castro, I. (2006). *Introdução à História do Português* (2ª edição revista e muito ampliada). Lisboa, Colibri.
- Cosme, J. et Varandas, J. (2009) *Memórias Paroquiais (1758), volume I [Abação-Alcaria]*, Casal de Cambra, Caleidoscópico.
- Galves, C.; Kato, M.; Roberts, I. (2019). *Português brasileiro. Uma segunda viagem diacrônica*. Campinas. Editora da Unicamp.
- Galves, C. M. C.; Britto, H. S.; Paixão de Sousa, M. C. (2005), «The Change in clitic placement from Classical to Modern European Portuguese: Results from the Tycho Brahe Corpus». *Journal of Portuguese Linguistics*, v. 4, p. 39-67.
- Gonçalves, M. F. (2003) *As ideias ortográficas em Portugal: de Madureira Feijó a Gonçalves Viana (1734-1911)*, Lisboa, Fundação Calouste Gulbenkian.
- Houaiss, A. (2001) *Grande Dicionário Houaiss da Língua Portuguesa*, Rio de Janeiro, Editora Objetiva.
- Kemmler, R. (2001), «Para uma História da Ortografia Portuguesa: o texto metaortográfico e a sua periodização do século XVI até à reforma ortográfica de 1911». *Lusorama* 47-48, pp. 128-319.
- Olival, F., Cameron, H.F, Farrica, F., Vieira, R. (2023) As Memórias Paroquiais do atual concelho de Vila Viçosa, Callipole, n.29, pp. 85-128.
- Pereira, B. (1697) *Prosodia in Dictionarium bilingue*, Évora, Évora, Tipografia da Academia, 7th edition.
- Reynaert, M. W. C., Hendrickx, I. H. E., & Marquilhas, R. (2012). «Historical spelling normalization. A comparison of two statistical methods: TICCL and VARD2». In F. Mambrini, M. Passarotti, & C. Sporleder (Eds.), *Proceedings of the Second Workshop on Annotation of Corpora for Research in the Humanities (ACRH-2)*, Lisbon, pp. 87-98. Edicoes Colibri.
- Roberts, I. et Kato, M. A. (1993). *Português brasileiro. Uma viagem diacrônica*. Editora Contexto.
- Santos, I., Olival, F., Sequeira, O., (2020), «Excavating the data pit: the Portuguese Parish Memories (1758) as a gold standard», in *DHandNLP 2020: Digital Humanities and Natural Language Processing: Proceedings of the Workshop on Digital Humanities and Natural Language Processing (DHandNLP 2020)* co-located with International Conference on the Computational Processing of Portuguese (PROPOR 2020). ed by M. José Finatto; Renata Vieira; Senja Pollak; Saturnino Luz, Évora, Vol. 2607.

Santos, J., Cameron, H.F., Olival, F., Farrica, F., Vieira, R. (2024) Named entity recognition specialised for portuguese 18th-century history research. In: *Proceedings of the 16th International Conference on Computational Processing of Portuguese*. pp. 117–126

Vieira, R., Olival, F., Cameron, H., Santos, J., Sequeira, O., Santos, I. (2021) Enriching the 1758 portuguese Parish Memories (Alentejo) with Named Entities. *Journal of Open Humanities Data* 7, 20

A sílaba e a composicionalidade em emakhuwa (P31): análise de empréstimos do português

Francelino Wilson¹, Vasco Magona², Felermino Ali³

¹Faculdade de Letras, Ciências Sociais e Humanidades – Universidade Púnguè
Chimoio, Moçambique

²Faculdade de Letras e Ciências Sociais – Universidade Eduardo Mondlane
Maputo, Moçambique

³Departamento de Engenharia da Computação – Universidade Lúrio
Pemba, Moçambique

fradwilson2@gmail.com, vasco.sande@uem.mz,
felermino.ali@unilurio.ac.mz

Abstract. *The syllabic model of Emakhuwa (P31) appears to differ from that of Portuguese, although they have in common the structure most attested in natural languages, i.e., CV. Based on this premise, it is interesting to carry out a contrastive analysis of the two languages, with the syllable as the scope of the study, to understand the compositionality in Emakhuwa. From the evaluation of unusual structures in the language, essentially formed by consonant and vowel sequences, in Portuguese nominal loans, it is concluded that emakhuwa is based on the canonical model of the CV syllable, although in specific situations the syllables of the V, N, VV and CVV types. The CV(V)C syllable format occurs only in some cases of structures borrowed from Portuguese.*

Resumo. *O(s) modelo(s) silábico(s) do emakhuwa (P31) parece diferir(em) da língua portuguesa, embora tenham em comum a estrutura mais atestada nas línguas naturais, i.e., CV. Partindo desta premissa, mostra-se interessante proceder a análise contrastiva das duas línguas, tendo como escopo do estudo a sílaba, para entender a composicionalidade em emakhuwa. Da avaliação de estruturas pouco comuns na língua, essencialmente formadas por sequências consonantais e vocálicas, em empréstimos nominais do português, conclui-se que o emakhuwa pauta pelo modelo canónico da sílaba CV, embora em situações pontuais possam ocorrer na língua outros tipos de sílabas, a exemplo de V, N, VV e CVV. Entretanto, o formato CV(V)C ocorre apenas em alguns casos de estruturas emprestadas do português.*

1. Introdução

A sílaba (σ) é um dos constituintes prosódicos que confere a composicionalidade às línguas naturais, a par do tom e do acento (Hyman & Mtenje, 1999; Mateus et al., 2003). O papel central deste elemento prosódico evidencia-se na cadeia de fala, disposta sob *continuum*, podendo abarcar mais de um segmento, situando-se, assim, acima deste (Andrade, 2020; Baylon & Fabre, 1979; Malmberg, 1993; Ngunga, 2014). Prova isso, a aquisição da linguagem pela criança, feita de combinação de sílabas, a partir da fase de

balbucio, para a formação das primeiras palavras, como *ba.ba* ou *pa.pa*, *ma.ma*; idem, a aprendizagem da leitura, que se mostra frutífera, para muitas crianças, pelo método silábico (ou da silabação) (Hachette, 1967 *apud* Genouvrier & Peytard, 1974:82).

O estudo da sílaba (ou da silabação) parece contribuir para: *a.* a compreensão da composicionalidade em emakhuwa (P31, na classificação de Guthrie, 1967-71); *b.* a fixação do padrão ortográfico de línguas com um histórico recente de gramaticalização ou normalização linguística, a exemplo do idioma objeto de análise; e *c.* a aplicação deste conhecimento em processamento da linguagem natural, objetivos que nos propomos alcançar com este trabalho. Na generalidade, pretende-se responder à questão segundo a qual: as palavras emprestadas do português (essencialmente constituídas por nomes) sujeitam-se às condições restritivas do emakhuwa, à luz dos modelos silábicos admissíveis na língua? De antemão, procura-se provar a hipótese segundo a qual só são aceites na língua de chegada as palavras que se sujeitam às condições de boa-formação nela impostas, a exemplo dos padrões silábicos reconhecidos.

A perspetiva teórica que se oferece à discussão que se segue obedece à Teoria da Otimidade (Abaurre, 1999; McCarthy, 1993; Prince & Smolensky, 1993; Yip, 1993), por se mostrar orientada para a identificação de restrições que regulam a boa-formação das representações. Do mesmo modo, a OT mostra-se eficaz para o estudo dos empréstimos, uma vez constituídos de palavras que se movem junto de condições de boa-formação e restrições impostas pela língua de chegada (Yip, 1993), no processo de escolha do candidato ótimo. Como tal, o modelo faz-se valer de propriedades como a violabilidade, a hierarquia, a inclusividade e o paralelismo (Abaurre, 1999; McCarthy, 1993; Prince & Smolensky, 1993), tendo como vantagens a economia descritiva, a universalidade e a uniformidade de análise (Abaurre, 1999; Collischonn & Schwindt, 2003).

A análise assenta em pouco mais de duzentos étimos nominais do português encontrados no emakhuwa, línguas em contato desde o século XV (Hagemeijer, 2016; Mota & Bacelar do Nascimento, 2001). A recolha do material empírico foi feita manualmente por sete tradutores, naturais da província de Nampula, em Moçambique, falantes natos da língua makhuwa, com formação em ensino de línguas de substrato Bantu. A cada participante do estudo foi designado um conjunto de textos contendo notícias do jornal VOA, redigidos em português, sobre vários domínios, incluindo saúde, sociedade, cultura, desporto e política. Uma vez traduzidos para emakhuwa, estes textos foram carregados na ferramenta ‘CAT Matecat’ (Federico et al., 2014) que segmenta os enunciados em pequenas unidades linguísticas, em documento respetivo. Após a conclusão da tradução, os profissionais foram instruídos a registrar as palavras emprestadas do português, obviamente sujeitas à adaptação fonético-fonológica, morfológica e semântica.

Emakhuwa é a língua amplamente falada em Moçambique, por ¼ de pouco mais de 30 milhões da população moçambicana com cinco ou mais anos de idade (INE, 2019, 2023), com o estatuto de língua materna e franca em algumas regiões do Norte do país (Lindonde, 2021). É igualmente falada no Malawi e na República Unida da Tanzânia (Bernardo, 2017; Kisseberth, 2003; Kisseberth & Cassimjee, 2009), idem em comunidades da África do Sul, Madagáscar, Ilhas Comores, Seychelles, Maurícias e Zanzibar (Kröger, 2005). Dada a extensão territorial por onde é falada, é marcada por contatos com outras línguas, dando lugar a línguas ou dialetos mistos, a exemplo de

echuwabo, ekoti e esangaji (Kisseberth, 2003; Kisseberth & Guérois, 2014). É também expectante que seja caracterizada por diversos fenómenos derivados do contato, entre os quais, os empréstimos lexicais.

Após o lançamento do escopo, da definição dos objetivos, da pergunta de partida, da hipótese, da perspetiva teórica, da caracterização dos dados e da rápida sociolinguística do emakhuwa, segue-se, em 2, um breve aporte teórico; a discussão de resultados e conclusões preliminares constam da seção 3.

2. Breve discussão contrastiva da sílaba em emakhuwa

É consensual afirmar-se que as unidades fonológicas significativas do emakhuwa incluem 10 vogais, entre breves (/a e i o u/) e longas (/aa ee ii oo uu/) (Centis, 1985; Kröger, 2006; Ngunga et al., 2022; Victorino, 1995), e um total de 25 consoantes contrastivas (Ngunga et al., 2022), entre elas, as semivogais (cf. tabela 1), sujeitas a fenómenos de modificação, cujos exemplos podem ser vistos na literatura de referência. O alongamento que se faz das vogais serve para expressar a duração que estas estão sujeitas, um aspeto contrastivo na língua (Mtenje, 2007; Ngunga, 2014; Victorino, 1995) e com implicações a nível da sílaba.

Tabela 1. Consoantes do emakhuwa

Modo/Lugar		Labial	Dental	Alveolar	Retroflexa	Palatal	Lábio-velar	Velar	Glotal
Oclusiva	plena	p	t		tt	c		k	
	aspirada	ph	th		tth	*(ch)		kh	
Fricativa		f v		s *(z)		x			h
Nasal		m		n		ny		*(ng)	
Lateral				l		ly			
Vibrante				r					
Semivogal						y	w		

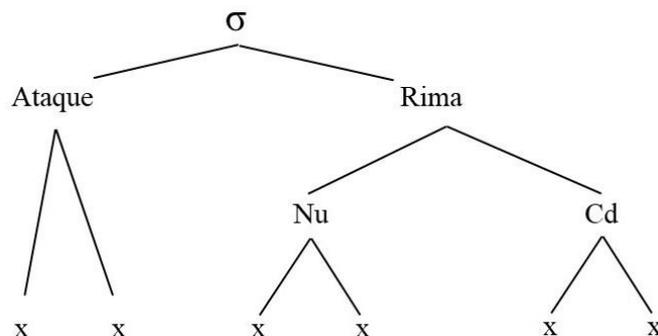
Fonte: (Ngunga et al., 2022:105)

* Assinala consoantes assumidas como marginais na língua

Para a conceituação da sílaba faz-se necessário conhecer algumas “regras” (ou construtos) de base, tais como a hierarquia prosódica, a condição de dissimilaridade e os princípios de licenciamento e sequenciamento prosódicos (Freitas & Santos, 2009; Mateus et al., 2003; Schwindt, 2024; Veloso, 2006), que constituem restrições para a boa-formação deste constituinte prosódico. Na generalidade, dada a limitação de espaço disponível nesta publicação, cada um destes princípios/regras concorre para à formação de unidades silábicas aceitáveis em cada uma das línguas naturais, cujo desvio sugere, no mínimo, estruturas emprestadas de outros idiomas nalguma fase de desenvolvimento da língua de chegada (Veloso, 2006; Vigário et al., 2006).

A natureza suprasegmental da sílaba, que possibilita a articulação de vários elementos na sua origem, faz dela um constituinte fonológico complexo (Andrade & D’Andrade, 2020; Blevins, 1995; Mateus & D’Andrade, 2000; Ngunga, 2014). Do conhecimento reunido até então, fazem parte da estrutura silábica o núcleo (Nu) e as margens pré-nuclear e pós-nuclear, formadas por consoante ou conjunto de consoantes simples ou modificada(s) (Kindell, 1981 *apud* Ngunga, 2014:90), também designadas, respetivamente, ataque (A) e coda (Cd) (cf. figura 1) (e. o., Blevins, 1995; van der Hulst & Smith, 1982).

Figura 1. Modelo hierárquico da estrutura da sílaba



Fonte: Adaptado de (Blevins, 1995; van der Hulst & Smith, 1982)

Desta forma, depreende-se que a sílaba é uma estrutura hierarquicamente organizada em constituintes silábicos que, de acordo com as propostas contemporâneas (Andrade & D’Andrade, 2020; Freitas & Santos, 2009; Schwindt, 2024), apresentam, no máximo, duas posições internas, designadamente, *i.* a sílaba, ramificada em Ataque e Rima; e *ii.* a Rima, ramificada em Núcleo e Coda. Entretanto, cada constituinte está associado a um mínimo de uma e a um máximo de duas posições rítmicas, quando ramificado, no nível do esqueleto (x). Deve-se acrescentar que, cada posição rítmica, no esqueleto (x), pode ou não estar associada a material segmental, i.e., diz-se vazio (\emptyset), quando não se encontra preenchido com material linguístico.

São hoje conhecidos vários modelos silábicos, que se atualizam de língua para língua, ocasionando pelo menos 10 possibilidades, na generalidade, de V a CVCCC (Blevins, 1995). Em português, por exemplo, são conhecidos 14 padrões silábicos (Andrade & D’Andrade, 2020); em inglês, 12 (Schwindt, 2024); em hua, apenas um (Blevins, 1995).

As sílabas em emakhuwa são, regra geral, abertas ou livres, i.e., terminadas em vogal, exceto nos casos em que elas são formadas de nasais silábicas (cf. (1)), não se prevendo a ocorrência de codas (Ngunga, 2014).

- (1) n(i)-lok-o [n.lo.ko] ‘tribo’
 n(i)-rapal-a [n.ra.pa.la] ‘pele’

Os dados em (1) mostram que, “em emakhuwa, quando uma consoante nasal [autónoma] é seguida de um som com função de consoante [ganha] estatuto de sílaba, isto é, [torna-se] uma nasal silábica” (Ngunga et al., 2022:107). Os casos em apreço são resultantes da elisão da vogal /i/ do prefixo *ni-* classe 5, do singular, cujo plural se obtém pela substituição do prefixo *ni-* por *ma-* classe 6 (e.g., *ni-lok-o* > *n’-lok-o* ‘tribo’ vs. *ma-lok-o* ‘tribos’; *ni-rapal-a* > *n’-rapal-a* ‘pele’ vs. *ma-rapal-a* ‘peles’).

Outro formato de sílaba atestado em emakhuwa é o formado por V, sem nenhuma margem, admissível unicamente em início de palavra (cf. (2)), a semelhança da nasal silábica descrita em linhas precedentes (Ngunga, 2014; van der Wal, 2009). É, de resto, um modelo silábico previsto nas línguas de substrato Bantu, não só, idem no português, onde tem a segunda maior frequência – 15,83% (Vigário et al., 2006) –, depois de CV – 46,36% (Vigário et al., 2006) –, o formato *default* em muitas línguas naturais, idem nas línguas Bantu, de que emakhuwa é parte integrante.

(7) n(i)-tthup-i [nt.thu.pi] ‘poeira’ (van der Wal, 2009:17)

Vistas as coisas sob este prisma, contrariamente aos sete tipos de sílabas propostos por van der Wal (2009), passamos a ter um quadro muito mais extenso de ocorrências silábicas em emakhuwa (cf. tabela 2).

Tabela 2. Proposta de modelos silábicos ocorrentes em emakhuwa

	Sílaba	exemplo
1	V	e-hop-a [e.ho.pa] ‘peixe’
2	N	n(i)-ter-e [n.te.re] ‘lábio’
3	CV	o-low-a [o.lo.wa] ‘pescar’
4	VV	oo-low-a [oo.lo.wa] ‘ele pescou’
5	CVV	o-khool-a [o.khoo.la] ‘moer’
6	CVN	e-mank-a [e.man.ka] ‘manga’
7	? CVC	mallim-u [mal.li.mu] ‘professor da madraça’
8	CVC _i	o-patt-a [o.pat.ta] ‘acender a luz’
9	VC _i	o-ttek-a [ot.te.ka] ‘obstruir, abrir guarda-chuva’
10	CVVC _i	weett-a [weet.ta] ‘andar’
11	NC _i	n(i)-tthup-i [nt.thu.pi] ‘poeira’

Fonte: os autores, de várias leituras

Apesar da multiplicidade de modelos silábicos, persistem diferenças notáveis entre o padrão silábico desta língua com o do português. Entre essas diferenças, podem ser referidas, no emakhuwa, diferentemente do português, a impossibilidade de formação de ataques ramificados, a ocorrência de nasais silábicas e a preferência por sílabas abertas (logo, a inexistência de codas). Sobre este último aspeto, em sílabas do tipo CVN, não parece N# tratar-se de coda, antes, porém, um autosegmento flutuante, passível de se associar à V precedente ou à C seguinte para lhe conferir a nasalidade (cf. Anderson, 1976; Júnior, 2022; Moraes & Wetzels, 1992).

3. A sílaba e a composicionalidade em étimos do português

Em Moçambique, falantes que têm uma língua de substrato como L1 usam palavras da sua língua materna Bantu para preencher as lacunas lexicais do português, em domínios como a fauna, a flora, as práticas culturais e/ou servem-se do léxico português no discurso corrente Bantu (Gonçalves, 2012; Machungo, 2022; Mendes, 2012). Diferente da cunhagem de termos, aquele recurso mostra-se economicamente viável, por ser menos dispendioso no labor terminológico, o que aumenta a sua recorrência pelos profissionais da tradução (português-Bantu).

Neste trabalho, que se prevê exploratório, a análise dos dados restringe-se à discussão do padrão silábico do emakhuwa, na condição de língua que define a aceitabilidade das construções provenientes da língua portuguesa. Na impossibilidade de se fazer uma abordagem exaustiva, dada a complexidade do fenómeno prosódico e escassez de espaço, circunscrevemo-nos aos modelos silábicos menos atestados em emakhuwa (cf. tabela 3). Deste modo, as categorias de análise são formadas por étimos nominais do português que contêm sílabas com estruturas envolvendo: *a.* sequência consonantal com grupo próprio, i.e., obstruinte + líquida (em diante, CC); *b.* sequência de obstruinte /ʃ/ + consoante (SC); *c.* sequência de obstruinte /ʃ/ + obstruinte diferente de /ʃ/ + líquida (SCC); *d.* sílaba fechada, i.e., sílaba terminada em coda (CVC); *e.*

sequência de consoantes pertencentes a sílabas distintas (C.C); e, por último, *f*. sequência vocálica (VV).

Tabela 3. Categorias de análise formadas por étimos nominais do português

Categorias de análise	Étimos do português
CC	empresa, príncipe, branca, Brasil, febre, abril, fábrica, outubro, tribunal, drone, África, integridade
SC	Moscovo, pesquisa, ativista, Israel
SCC	Austral, ministro, administração
CVC	Brasil, tribunal, parlamento, partido, pesquisador, palma
C.C	Arquipélago, parlamento, administração, partido, Israel
VV	outubro, Austral, Israel

Fonte: os autores

Apresentadas as categorias de análise, segue-se uma rápida discussão à volta da entrada dos étimos nominais do português no emakhuwa, à luz da OT. Dadas às limitações já referidas, dá-se um caso por cada categoria, conforme os candidatos a *output* e as possíveis restrições, partindo do modelo CC (cf. (8)).

(8)

/êprezɛ/	Não ramificação ^{ataque}	Não coda	Harmonia vocálica
☞ a. ê.pe.re.zɛ			
b. ê.pi.re.zɛ			*
c. ê.pre.zɛ	*!		*

Em (8), para a formação do candidato ótimo, não basta que, entre o grupo próprio, se insira uma vogal epentética para se evitar a sequência CC, pouco produtiva nas línguas Bantu. Mas, a vogal inserida tem que se sujeitar à regra de harmonia vocálica (cf. Alves, 2011; Ngunga, 2014), o que faz com que seja preferida /e/, similar às vogais das sílabas vizinhas, em relação a /i/, apontada como canónica no dialeto brasileiro do português (Mateus *et al.*, 2003). A construção em (9), por sua vez, parece sugerir um caso contrário ao aqui descrito.

(9)

/muʃkovu/	Não ramificação ^{ataque}	Não coda	Epêntese
☞ a. muʃ.ko.vu		*	
b. mu.ʃu.ko.vu			*!
c. mu.ʃi.ko.vu			*!

Nesta ocorrência, a inserção de vogal epentética (quer /u/, como /i/) não parece ser a solução para a evitação de encontros SC. Antes, porém, o alongamento da vogal da sílaba portadora de coda se afigura a solução. Esta situação tem como consequência a introdução na língua de um modelo silábico pouco atestado, pelo menos, em emakhuwa, ainda que admissível em empréstimos do português, língua com padrão silábico CVC. Em (10), analisa-se um étimo contendo estruturas do tipo SCC.

(10)

/miniʃtru/	Não ramificação ^{ataque}	Não coda	Epêntese
☞ a. mi.ni.ʃi.tu.ru			*
b. mi.ni.ʃ.tu.ru		*!	*
c. mi.ni.ʃ.tru	*!	*!	

Para este caso, as violações gravosas são as impostas pelas restrições de formação de ataques ramificados e codas silábicas, não preferenciais em emakhuwa. Em consequência, emerge o modelo silábico CV, como o mais atestado, idem na construção em (11), cujo alongamento da vogal da penúltima sílaba, propensa ao alongamento não fonémico, dá lugar ao modelo CVV. Aqui, a vogal epentética convocada para a evitação da formação de codas silábicas é /u/, contrariamente ao esperado pela aplicação da regra de harmonia vocálica.

(11)

/pałmɛ/	Não ramificação ^{ataque}	Não coda	Epêntese
☞ a. paa.lu.mɛ			*
b. pa.la.mɛ			*
c. pał.mɛ		*!	

Em (12), os candidatos a *output* que se mostram violarem restrições não toleráveis na língua de chegada são os que apresentam ataque complexo (cf. (12b)) e coda silábica (cf. (12c)). A inserção da vogal /i/ entre a sequência de consoantes pertencentes a sílabas distintas, i.e., com divisão *heterossilábica* (C.C), pelo contrário, mostra-se uma restrição violável que, por sinal, concorre para a formação de sílabas autónomas (i.e., *heterossilábicas*) de tipo CV.CV, de igual estatuto vocálico. A inserção deste material segmental evita a presença na língua de codas e sequências de CC em posição de ataque. Outra consequência é o aparecimento do modelo silábico do tipo V, em emakhuwa, de resto um figurino atestado na língua.

(12)

/ɛrkiɲɛlɛgu/	Não ramificação ^{ataque}	Não coda	Epêntese
☞ a. a.ri.ki.pee.la.ku			*
b. a.rki.pee.la.ku	*!		
c. ar.ki.pee.la.ku		*!	

Como em (12), a presença do modelo silábico de natureza V, em (13), é decorrente da monotongação da primeira sílaba do *input*. Aqui, tanto a inserção de material segmental (cf. (13b)), como a formação de um núcleo ramificado (cf. (13c)), revelam-se estratégias não viáveis em emakhuwa, no contexto em discussão.

(13)

/owtubru/	Não ramificação ^{Nu}	Não coda	Epêntese
☞ a. u.tu.pu.ru			*
b. o.wu.tu.pu.ru			*
c. ow.tu.bru	*!		

Em conclusão, embora preliminar, o emakhuwa parece pautar pelo modelo canónico da sílaba CV, podendo, em situações pontuais, admitir outros formatos, como V, N, VV e CVV. A hipótese levantada na introdução fica grandemente provada, salvo em casos de formação de sílabas do tipo CV(V)C (cf. (9)), forma emprestada do português. Aparentes situações de ocorrência de codas são desfeitas, quer pela atribuição de valor de mora ou autosegmento nasal às consoantes nessa posição, quer por estratégias conducentes à ressilabificação, a exemplo da inserção de material segmental depois de C.

Referências

- Abaurre, M. B. M. (1999). Teoria da otimalidade e fonologias derivacionais. *Cadernos de Estudos Linguísticos*, 37, 69–80.
- Alves, M. M. (2011). Harmonia vocálica e redução vocálica à luz da teoria da otimalidade. *Anais Do SILEL*, 2(2), 1–19. <http://www.ileel.ufu.br/anaisdosilel/pt/arquivos/silel2011/2512.pdf>
- Anderson, S. R. (1976). Nasal Consonants and the Internal Structure of Segments. *Language*, 52(2), 326–344. <https://doi.org/10.2307/412563>
- Andrade, A. (2020). Vocalismo. In E. B. P. Raposo, M. F. Bacelar do Nascimento, M. A. C. da Mota, L. Segura, A. Mendes, & A. Andrade (Eds.), *Gramática do Português, Vol. III* (pp. 3241–3330). Fundação Calouste Gulbenkian.
- Andrade, A., & D'Andrade, E. (2020). Sílabas. In E. B. P. Raposo, M. F. Bacelar do Nascimento, M. A. C. da Mota, L. Segura, A. Mendes, & A. Andrade (Eds.), *Gramática do Português, Vol. III* (pp. 3369–3396). Fundação Calouste Gulbenkian.
- Baylon, C., & Fabre, P. (1979). *Iniciação à Linguística*. Livraria Almedina.
- Bernardo, M. (2017). *Análise Fonológica da Estrutura Verbal do Passado Recente em Emakhuwa*. (Dissertação de Mestrado). Universidade Eduardo Mondlane.
- Blevins, J. (1995). The Syllable in Phonological Theory. In J. A. Goldsmith (Ed.), *The Handbook of Phonological Theory* (pp. 206–244). Blackwell.
- Botne, R. (2019). Chimpoto N14. In M. Van de Velde, K. Bostoen, D. Nurse, & G. Philippson (Eds.), *The Bantu Languages* (2nd Editio, pp. 692–732). Routledge.
- Centis, G. (1985). *Método Macua*. Centro Catequético Paulo VI.
- Collischonn, G., & Schwindt, L. C. (2003). Teoria da Otimalidade em Fonologia: discutindo conceitos. In D. Da Hora & G. Collischonn (Eds.), *Teoria Linguística: fonologia e outros temas* (pp. 17–50). Editora Universitária/UFPB.
- Federico, M., Bertoldi, N., Cettolo, M., Negri, M., Turchi, M., Trombetti, M., Cattelan, A., Farina, A., Lupinetti, D., Martines, A., Massidda, A., Schwenk, H., Barrault, L., Blain, F., Koehn, P., Buck, C., & Germann, U. (2014). The MateCat Tool. *COLING 2014 - 25th International Conference on Computational Linguistics, Proceedings of the Conference System Demonstrations*, 7(287688), 129–132.
- Freitas, M. J., & Santos, A. L. (2009). *Contar (histórias de) sílabas: descrição e implicações para o Ensino do Português como Língua Materna* (2ª Edição). Edições Calibri e Associação de Professores.
- Genouvrier, E., & Peytard, J. (1974). *Linguística e Ensino do Português*. Livraria Almedina.
- Gonçalves, P. (2012). Contacto de línguas em Moçambique: algumas reflexões sobre o papel das línguas bantu na formação de um novo léxico do português. In T. Lobo, Z. Carneiro, J. Soledade, A. Almeida, & S. Ribeiro (Eds.), *ROSAE - Linguística Histórica, História das Línguas e Outras Histórias* (pp. 401–405). EDUFBA.
- Guthrie, M. (1967). *Comparative Bantu: an introduction to the comparative linguistics*

and prehistory of the Bantu languages. Gregg International.

- Hagemeyer, T. (2016). O português em contacto em África. In A. M. Martins & E. Carrilho (Eds.), *Manual de linguística portuguesa* (pp. 43–67). De Gruyter.
- Hyman, L. M. (1985). A Theory of Phonological Weight. In *Language* (Vol. 62, Issue 3). Foris. <https://doi.org/10.2307/415484>
- Hyman, L. M., & Katamba, F. X. (1999). The syllable in Luganda phonology and morphology. In H. van der Hulst & N. Ritter (Eds.), *The Syllable: Views and Facts* (pp. 349–416). De Gruyter Mouton. <https://doi.org/10.1515/9783110806793.349>
- Hyman, L. M., & Mtenje, A. (1999). Prosodic Morphology and tone: the case of Chichewa. In R. Kager, H. van der Hulst, & W. Zonneveld (Eds.), *The Prosody-Morphology Interface* (pp. 90–133). Cambridge University Press. <https://doi.org/10.1017/cbo9780511627729.005>
- INE. (2019). *IV Recenseamento Geral da População e Habitação 2017: resultados definitivos Moçambique*. Instituto Nacional de Estatística.
- INE. (2023). *Padrão Linguístico em Moçambique*. Instituto Nacional de Estatística.
- Júnior, R. B. (2022). Uma abordagem fonético-fonológica das vogais nasais do Português de Moçambique. In H. N. Dias & I. E. Taela (Eds.), *Português Moçambicano I* (pp. 118–126). Alcance Editores.
- Kisseberth, C. W. (2003). Makhuwa (P30). In D. Nurse & G. Philippson (Eds.), *The Bantu Languages* (pp. 546–565). Routledge.
- Kisseberth, C. W., & Cassimjee, F. (2009). *The Emakhuwa Lexicon Exemplified*.
- Kisseberth, C. W., & Guérois, R. (2014). Melodic H tones in Emakhuwa and Ecuwabo verbs. *Africana Linguistica*, 20, 181–205. <https://doi.org/10.3406/aflin.2014.1030>
- Kröger, O. (2005). Report on a Survey of Coastal Makua Dialects. In *SIL International*. <http://www.sil.org/silesr/abstract.asp?ref=2005-020>
- Kröger, O. (2006). *Algumas notas gramaticais sobre a língua Emakhuwa*. Sociedade Internacional de Linguística.
- Lindonde, L. M. (2021). A Questão de Escolha Linguística em Ambientes Domésticos num Contexto Multilingue de Moçambique. *Alfa*, 65, 1–17. <https://doi.org/https://doi.org/10.1590/1981-5794-e12448>
- Machungo, I. (2022). Empréstimos lexicais no Português de Moçambique. In E. A. C. Nhampoca & L. C. Ponso (Eds.), *Pesquisa e ensino em línguas moçambicanas: um tributo a Bento Siteo* (pp. 100–113). Gala-Gala Edições.
- Malmberg, B. (1993). A fonética: teoria e aplicações. *Cadernos de Estudos Linguísticos*, 25, 7–24.
- Mateus, M. H. M., & D'Andrade, E. (2000). *The Phonology of Portuguese*. Oxford University Press.
- Mateus, M. H. M., Frota, S., & Vigário, M. (2003). Prosódia. In M. H. M. Mateus, A. M. Brito, I. Duarte, I. H. Faria, S. Frota, G. Matos, F. Oliveira, M. Vigário, & A. Villalva (Eds.), *Gramática da Língua Portuguesa* (7ª Edição, pp. 1035–1076). Caminho.

- McCarthy, J. J. (1993). A Case of Surface Constraint Violation. *Canadian Journal of Linguistics/Revue Canadienne de Linguistique*, 38(2), 169–195. <https://doi.org/10.1017/s0008413100014730>
- Mendes, I. (2012). Empréstimos no Português de Moçambique. *DEBATES*, 12, 1–17.
- Moraes, J. A., & Wetzels, W. L. (1992). Sobre a duração dos segmentos vocálicos nasais e nasalizados em português. Um exercício de fonologia experimental. *Cadernos de Estudos Linguísticos*, 23, 153–166.
- Mota, M. A., & Bacelar do Nascimento, M. F. (2001). Le portugais dans ses variétés. *Revue Belge de Philologie et d'histoire*, 79(3), 931–952. <https://doi.org/10.3406/rbph.2001.4554>
- Mtenje, A. (2007). On Recent Trends in Phonology: Vowel Sequences in Bantu Languages. *SOAS Working Papers in Linguistics*, 15, 33–48.
- Ngunga, A. (2014). *Introdução à Linguística Bantu* (2ª Edição). Imprensa Universitária/UEM.
- Ngunga, A., Manuel, C., Langa, D., Machungo, I., & Câmara, C. L. da. (2022). *Padronização da Ortografia de Línguas Moçambicanas: Relatório do IV Seminário*. Imprensa Universitária/UEM.
- Prata, A. P. (1990). *Dicionário Macua-Portugues*. Instituto de Investigação Tropical.
- Prince, A., & Smolensky, P. (1993). *Optimality Theory: Constraint Interaction in Generative Grammar*. Rutgers University. <https://doi.org/10.1002/9780470756171.ch1>
- Schwindt, L. (2024). Sílabas. In *Speech Sciences Entries*. Speech Prosody Studies Group. <https://gepf.falar.org/entries/55>
- van der Hulst, H., & Smith, N. (1982). An Overview of Autosegmental and Metrical Phonology. In H. van der Hulst & N. Smith (Eds.), *The structure of Phonological Representations. Part 1* (pp. 1–45). Foris. <https://doi.org/10.1515/9783112328088-002>
- van der Wal, J. (2009). *Word order and information structure in Makhuwa-Enahara*. (Ph.D Thesis). Leiden University.
- Veloso, J. (2006). Reavaliando o Estatuto Silábico das Sequências Obstruinte + Lateral em Português Europeu. *D.E.L.T.A: Documentação de Estudos Em Linguística Teórica e Aplicada*, 22(1), 127–158.
- Victorino, A. (1995). *Estudo Comparativo Fonológico das Variantes do Emakhuwa: Implicações Ortográficas*. (Monografia de Licenciatura). Faculdade de Letras da Universidade Eduardo Mondlane.
- Vigário, M., Martins, F., & Frota, S. (2006). A Ferramenta FreP e a frequência de tipos silábicos e classes de segmentos no Português. In J. Barbosa & F. Oliveira (Eds.), *Textos seleccionados do XXI Encontro Nacional da Associação Portuguesa de Linguística (Porto 2005)* (pp. 675–687). Colibri.
- Yip, M. (1993). Cantonese loanword phonology and Optimality Theory. *Journal of East Asian Linguistics*, 2(3), 261–291. <https://doi.org/10.1007/BF01739135>

Linguistic and emotional dynamics in satirical vs. real news: a psycholinguistic analysis

Gabriela Wick-Pedro¹, Roney Lira de Sales Santos², Oto Araújo Vale³

¹Instituto de Ciência, Tecnologia e Inovação Ciência e Tecnologia (IBICT)
Brasília – DF – Brazil

²Instituto de Ciência, Tecnologia e Inovação – Universidade Federal da Bahia (UFBA)
Camaçari – BA – Brazil

³Programa de Pós-graduação em Linguística – Universidade Federal de São Carlos (UFSCar)
São Carlos – SP – Brazil

`gabrielawick@ibict.br, roneysantos@ufba.br, otovale@ufscar.br`

Abstract. *This study compares the psycholinguistic differences between satirical and real news using data from LIWC (Linguistic Inquiry and Word Count). We found that satirical news utilizes a broader range of emotional and rhetorical resources, often exaggerating or subverting reality, while real news maintains a more factual and objective tone. This highlights the critical and humorous role of satire in social communication. Furthermore, the research advances the field of NLP by improving satire detection through a psycholinguistic lens, contributing to the development of algorithms that effectively differentiate satirical news from fake news and help combat misinformation.*

1. Introduction

The complexity of satire arises from its potential to "reinterpret" factual news, thereby blending real information with fictional elements to convey a subjective message. This creates a significant challenge in distinguishing satirical content from genuine news, especially as satire may incorporate authentic events and details, causing overlap between factual and fictional elements [Burfoot and Baldwin 2009]. The ability to automatically identify satirical news is crucial for developing effective tools to combat misinformation and deceptive content.

Satirical news and fake news are interconnected phenomena that emerge in the context of contemporary communication, particularly on digital platforms. While fake news, or false news, consists of misleading or fictitious information presented as true, with the deliberate intention of deceiving the public for purposes of propaganda, financial gain, or confusion [Rubin et al. 2015], [Wardle and Derakhshan 2017], [Tandoc Jr et al. 2018], satirical news employs humor and irony as tools to provoke critical reflection on social and political issues. This difference in intent is fundamental to understanding how each type of content is perceived and interpreted by readers.

However, the humorous nature of satirical news can easily be lost, leading to confusion between what is real and what is fictional. The widespread sharing of satirical content on social media often obscures the lines between satire and reality, resulting in the inadvertent dissemination of false information when the satire is not correctly identified. This phenomenon highlights the challenge of reader interpretation, especially in

digital environments where the context and intent of the message may not be immediately clear. As noted by [Friggeri et al. 2014], 'the rapid spread of satirical content on social media often blurs the lines between fact and fiction, contributing to the confusion experienced by audiences.' Thus, analyzing the differences between these two types of content is crucial for developing effective strategies for identifying and mitigating misinformation in the digital age.

To address these challenges, this study explores psycholinguistic relationships using data extracted from the LIWC (Linguistic Inquiry and Word Count) [Pennebaker 2001], [Balage Filho et al. 2013]. The LIWC is a computational text analysis tool designed to read each word in a text (or group of words) and match it with reference dictionaries loaded into the program's memory. [Yang et al. 2017] emphasized the importance of psychological characteristics in this context, noting that real news tends to be more conservative, while satirical news is often more aggressive. [Salas-Zárate et al. 2014] identified five significant psychological characteristics for detecting satirical news: social processes, affective processes, positive emotions, cognitive processes, and certainty.

The present research utilizes LIWC data to explore the psycholinguistic differences between satirical and real news content. A corpus was constructed with satirical news articles sourced from the Sensacionalista website, and corresponding real news articles were manually selected from reliable sources. The analysis focuses on key psycholinguistic features within the LIWC framework, examining categories such as family, affect, anger, and anxiety, among others, to differentiate between real and satirical news. By identifying these features, the research seeks to enhance the understanding of satire and improve automated methods for distinguishing satirical content from genuine news.

The article is organized as follows: Section 2 describes the corpus used in the analysis, outlining the collection process and characteristics of both satirical and real news texts. Section 3 presents the main psycholinguistic features extracted by LIWC and discusses their implications. Section 4 provides a detailed analysis of the results, highlighting the key differences between the two types of news. Finally, Section 5 offers the conclusions and discusses potential directions for future research.

2. Corpus

Considering the criteria proposed by [Rubin et al. 2015] for a fake news *corpus*, which emphasizes the importance of aligning fake news with true news to validate linguistic patterns through the identification of positive and negative instances, a *corpus* was created consisting of 300 news articles in the political domain, with 150 satirical news articles and 150 true news articles. The satirical news articles were automatically extracted from the Sensacionalista website, an online news outlet that uses the journalistic format to humorously address various current topics in Brazilian politics and entertainment. For the true news articles, the collection was conducted manually, initially identifying keywords from the satirical news articles and subsequently performing a manual search for each true news article equivalent to the satirical one. Table 1 details the characteristics of the *corpus*.

The data used to describe the corpus characteristics mentioned above were gen-

Table 1. Characteristics of the *corpus*

News	Tokens	Types	Sentences
Satirical	22.963	4.843	1.212
Real	107.133	11.304	5.721

erated using NLTK¹ and spaCy², two open-source libraries that offer valuable tools for natural language processing in Python. These libraries are widely employed both in production environments and in the development of applications capable of processing and understanding large volumes of text. To avoid losing information during analysis, the decision was made not to balance the corpus, as the number of words, sentences, and lexical diversity may serve as important descriptive features of this type of content. However, in future iterations of this work, a balance between the number of real and satirical texts will be sought, ensuring the integrity of the linguistic data and the representativeness of the categories analyzed.”

3. Features Extracted by LIWC

Each category represented in the LIWC reflects a specific aspect of language that can illustrate different dimensions of human experience and communication. For example, categories such as ”affect” and ”anger” help in understanding how emotions are expressed and manipulated in texts, while categories such as ”discrepancy” and ”certainty” reveal how texts address and represent controversial issues or absolute certainties.

The analysis of percentage differences between the two news categories indicates patterns that highlight how satirical news often amplifies or subverts aspects of reality to provoke reflection, humor, or social critique. In contrast, real news tends to maintain a more direct and factual approach, reflecting the informative and objective nature of traditional journalism.

Table 2 presents data extracted from the LIWC dictionary related to various psycholinguistic features in real and satirical news. The displayed percentages reflect the relative frequency of terms associated with emotions, mental states, and specific topics in both types of news. This analysis provides insights into how the language used in satirical texts differs from that in conventional news, highlighting the rhetorical and emotional strategies employed to engage the reader and provoke critical reflection.

Table tab:emoliwc presents the frequency of emotions identified in real and satirical news, as categorized by the LIWC dictionary. This table illustrates the distribution of words associated with three emotional categories: positive, negative, and neutral. The provided information enables a comparative analysis of the use of emotional terms in each type of news.

4. Analysis of results

The analysis of the data reveals significant differences in the psycholinguistic relationships between satirical and real news.

¹Available at: <https://www.nltk.org/index.html>.

²Available at: <https://spacy.io/>.

Table 2. Psycholinguistic relationships from LIWC extracted data

Features	Real	Satirical
achieve	5,70907483	5,80581862
anger	1,27298184	1,49304667
anx	0,59711395	0,93469965
cause	4,75479901	5,04649774
certain	2,9125677	3,15672724
death	0,33252275	0,39672383
discrep	5,1703802	5,81400334
family	0,595271	0,51880118
feel	1,65985041	1,98788499
friend	0,55472623	0,58740299
hear	1,80262334	2,23103831
home	0,91558896	0,95128402
inhib	6,05080874	6,24946677
insight	5,80394974	5,86980633
leisure	1,306142	1,98788499
money	2,86835415	2,84105452
relig	0,81242401	0,93848648
sad	0,71478575	0,86170122
see	1,17442246	1,04086682
swear	7,02166464	7,07277536
tentat	7,36431966	8,51036601
work	4,05198777	3,50652675

Table 3. Emotion scores of news articles based on the LIWC dictionary

Emotions	Real	Satirical
positive	0.109	0.392
negative	0.024	0.174
neutral	0.003	0,072

Satirical news employs terms related to achievements (5.8%) more frequently than real news (5.7%), suggesting a greater emphasis on exaggerating or highlighting accomplishments. Additionally, 5.046% of words in satirical news and 4.7% in real news fall under the category of “cause”.

The presence of words related to affect is higher in satirical news (3.9%) compared to real news (2.7%), indicating an amplification of emotional expressions to create contrast or emphasize absurdity.

The frequency of words associated with anger is greater in satirical news (1.4%)

than in real news (1.2%), which may reflect a tendency to underscore feelings of frustration or indignation.

Satirical news also shows higher levels of words related to anxiety (0.9%) compared to real news (0.5%), possibly to exaggerate concerns.

The use of terms related to “work” is lower in satirical news (3.5%) than in real news (4.0%), indicating a reduced focus on work-related issues. It is observed that satirical texts have higher levels of words related to attempts (8.5%) and certainty (3.1%), but lower levels of words related to work (3.5%). This may suggest that, due to their political domain, real news covers more topics related to political or legal work, while satirical news may be more focused on creating a sense of meaning and expressing possibility and uncertainty. As noted by [Salas-Zárate et al. 2017], the higher levels of certainty words (such as “never” or “always”) in satirical news (3.1%) compared to real news (2.9%) may be associated with hyperbole.

The discrepancy is more pronounced in satirical news (5.8%) than in real news (5.1%), suggesting a greater engagement in contesting or ridiculing ideas. Mentions of family are slightly less frequent in satirical news (0.51%) compared to real news (0.59%), possibly due to a greater focus on social or political issues.

Satirical news exhibits a higher frequency of words related to feelings (1.9%) compared to real news (1.6%), indicating a more intense exploration of emotions. The difference in mentions of friendships is minimal between the two types of news, suggesting a similar approach to interpersonal relationships, though satire may include critical nuances.

Words associated with “hearing” are more frequent in satirical news (2.2%) than in real news (1.8%), potentially highlighting aspects of communication or its absence.

Mentions of home are slightly more common in satirical news (0.95%) than in real news (0.91%), with no significant difference in the treatment of the theme.

Inhibition is somewhat more frequent in satirical news (6.2%) compared to real news (6.0%), which may reflect a more intense exploration of limitations. The presence of insights is slightly higher in satirical news (5.86%) compared to real news (5.80%), suggesting a more frequent use of insights to reinforce arguments or jokes.

The higher presence of terms related to leisure in satirical news (1.9%) may reflect a greater focus on trivial or everyday aspects, possibly to critique consumer society. Mentions of money are similar in both types of news, indicating that financial issues are equally relevant, although satire may address them critically.

Religion is mentioned more frequently in satirical news (0.9%) than in real news (0.8%), possibly to explore themes of faith or morality in a critical manner. Sadness is more prevalent in satirical news (0.8%) than in real news (0.7%), indicating a more frequent use of emotional themes to create impact.

Mentions of “seeing” are slightly less common in satirical news (1.0%) compared to real news (1.1%), suggesting a lower focus on direct visual observations. The frequency of swear words is similar in both types of news, suggesting that strong language is a common feature, with satire utilizing it more strategically.

Attempts are more common in satirical news (8.5%) than in real news (7.3%), reflecting a greater use of expressions of uncertainty or speculation. It is also noted that there are no completely divergent categories between real and satirical news. Satirical news features higher levels of words related to attempts, causes, and certainty, but lower levels of words related to work. This may indicate that, due to its political domain, real news covers more topics related to political or legal work, while satirical news may be more associated with creating a sense of meaning and expressing possibility and uncertainty. Additionally, as [Salas-Zárate et al. 2017] point out, the higher levels of certainty words in satirical news (3.1%) compared to real news (2.9%) may be associated with hyperbole. It is also noteworthy that 8.510% of words in satirical news and 7.3% of words in real news fall under the category of attempts, indicating that satirical texts make more use of modalizations, producing a sense of meaning and expressing a possibility and uncertainty.

In addition to the evident presence of affective terms in satirical news, validating that satirical texts are more emotional, the data shows that satirical news employs significantly more positive (0.392) and negative (0.174) words compared to real news. Thus, it can be asserted that the use of emotions (positive, negative, and neutral) is more pronounced in satirical news. It is interesting to consider that the use of emotion may be related to the fact that (1) real news tends to use more impartial language compared to satirical news and (2) the excessive use of positive words may, in fact, reflect the use of resources such as irony, which employs words or sentences with opposite polarity to the literal meaning, as also noted by [Salas-Zárate et al. 2017]. Additionally, the data show that satirical news contains a significantly higher proportion of words associated with positive (0.392) and negative (0.174) emotions compared to real news, which present 0.10% and 0.02%, respectively. Furthermore, the frequency of neutral words is also higher in satirical news (0.07%) compared to real news (0.003%). This difference in proportions may reflect the more emotional and expressive style of satirical news in contrast to the more impartial and factual approach of real news. Analyzing these data can provide insights into how different news genres use emotional language to create impact and engage readers.

5. Conclusions

The analysis of the data reveals striking differences in the psycholinguistic characteristics between satirical and real news. Satirical news employs a broader array of linguistic and emotional resources to create impact and convey social critique. The higher frequency of terms associated with achievements, affect, and attempts, as well as the more intense use of positive and negative emotions, suggests that satire employs exaggeration and subversion of reality to provoke reflection or humor. The pronounced presence of words related to uncertainty, certainty, and discrepancy reflects the nature of satire as a tool for challenging perceptions and questioning social norms. These differences indicate that satirical news not only highlights and amplifies aspects of reality but also utilizes linguistic resources to critique and satirize, offering a perspective that is often more emotive and critical compared to the more impartial approach of real news. Thus, satire emerges as an effective means for exploring and commenting on social and political issues, employing a wider range of discursive strategies to engage the audience and stimulate critical reflection.

Acknowledgements

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. The project was also supported by 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.

References

- Balage Filho, P. P., Pardo, T. A. S., and Aluísio, S. M. (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*, pages 215–219.
- Burfoot, C. and Baldwin, T. (2009). Automatic satire detection: Are you having a laugh? In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 161–164, Suntec, Singapore. Association for Computational Linguistics.
- Friggeri, A., Adamic, L., Eckles, D., and Cheng, J. (2014). Rumor cascades. In *proceedings of the international AAAI conference on web and social media*, volume 8, pages 101–110.
- Pennebaker, J. W. (2001). Linguistic inquiry and word count: Liwc 2001.
- Rubin, V. L., Chen, Y., and Conroy, N. K. (2015). Deception detection for news: three types of fakes. *Proceedings of the Association for Information Science and Technology*, 52(1):1–4.
- Salas-Zárate, M. d. P., López-López, E., Valencia-García, R., Aussenac-Gilles, N., Almela, Á., and Alor-Hernández, G. (2014). A study on liwc categories for opinion mining in spanish reviews. *Journal of Information Science*, 40(6):749–760.
- Salas-Zárate, M. d. P., Paredes-Valverde, M. A., Rodriguez-García, M. Á., Valencia-García, R., and Alor-Hernández, G. (2017). Automatic detection of satire in twitter: A psycholinguistic-based approach. *Knowledge-Based Systems*, 128:20–33.
- Tandoc Jr, E. C., Lim, Z. W., and Ling, R. (2018). Defining “fake news” a typology of scholarly definitions. *Digital journalism*, 6(2):137–153.
- Wardle, C. and Derakhshan, H. (2017). Information disorder: Toward an interdisciplinary framework for research and policy making. *Council of Europe*, 27.
- Yang, F., Mukherjee, A., and Dragut, E. (2017). Satirical news detection and analysis using attention mechanism and linguistic features. *arXiv preprint arXiv:1709.01189*.

Complementos de eco de adjetivos com completiva-sujeito em português do Brasil

Ryan Saldanha Martinez¹, Jorge Baptista², Oto Vale¹

¹Universidade Federal de São Carlos (UFSCar), Brasil

²Universidade do Algarve, Portugal / INESC-ID Lisboa, Portugal

{ryan.saldanha.martinez@gmail.com, jrbaptis@ualg.pt, otovale@ufscar.br}

Abstract. *This article discusses adjective complements that are coreferential with arguments of their subject clauses in Brazilian Portuguese. We propose that these complements are more economically conceived as increments, which we refer to as “echo complements”. We identify and syntactically describe three types of echo complements: da parte de N^0 (of N^0), para com N^1 (with N^1), and para N^0 (for N^0).*

Resumo. *Este artigo discute complementos de adjetivo correferentes a argumentos de sua completiva-sujeito em PB. Propomos que esses complementos podem ser mais economicamente caracterizados como incrementos, denominando-os “complementos de eco”. Identificamos e descrevemos as propriedades sintáticas de três tipos de complementos de eco: da parte de N^0 , para com N^1 e para N^0 .*

1. Introdução

Este artigo tem como objeto a análise sintática de um tipo complemento de adjetivo correferente a um item previamente mencionado ou pressuposto na completiva-sujeito desse adjetivo, como em (1)¹.

- (1) a. *Que o João tenha feito isso é honroso*
 b. *Que o João_i tenha feito isso é honroso da parte dele_i*

Adjetivos como *honroso* apresentam uma construção com completiva-sujeito (1a) e podem ter acrescido um complemento do tipo *da parte de N* (1b), sendo *N* correferente ao sujeito da completiva-sujeito. Complementos que apresentam essas propriedades de correferência são conhecidos como *complementos de eco* [Guillet and Leclère 1981, pp. 116–117], [Vivès 1982], [Baptista 2005, pp. 157–162], [Baptista and Mamede 2013].

Neste artigo, propomo-nos a descrever a sintaxe dos diferentes tipos de complementos de eco de adjetivos que se constroem com completiva-sujeito. Essa descrição se inscreve no quadro da construção de uma *gramática mínima* (do ing. *least grammar*, [Harris 1991, p. 4]), isto é, uma gramática que reduza ao mínimo as redundâncias na

¹Nos exemplos, os índices *i* indicam uma relação de correferência entre duas palavras; utiliza-se, adicionalmente, *j* quando dois índices são necessários.

descrição de uma dada língua, sendo, contudo, capaz de gerar todas as suas frases gramaticais.

Nessa teoria, a ideia de uma transformação incremental que adiciona redundância linguística a uma frase remonta a [Harris 1970, p. 501] e [Harris 1982, pp. 14–15]. Harris a caracteriza como o *inverso* de uma operação de redução por redundância, como as que ocorrem entre elementos unidos por conjunção que se repetem, tal como “o João” e “fez” na frase *O João fez isso e o João fez aquilo*, reduzidos a zero em *O João fez isso e aquilo*.

Assim, diante de alternâncias como (1), optar-se-á por uma entre duas possibilidades teóricas: (i) tomar (1a) como forma de base e caracterizar (1b) como resultado de uma transformação incremental, visão adotada neste artigo; ou (ii) tomar (1b) como forma de base e caracterizar (1a) como resultado de uma redução do elemento redundante, solução adotada, entre outros, em [Casteleiro 1981, Ranchhod 1985].

Conceber formas do tipo (1b) como incrementos a formas do tipo (1a) provê uma maneira econômica de se formalizar tais alternâncias. Os complementos de eco não carregam informação nova; apenas reiteram parte da informação que consta na forma da base, nomeadamente o sujeito ou complemento do predicado que ocupa a posição de sujeito. Ora, se considerarmos que as frases de base devem conter toda a informação da língua, um elemento opcional dentro da frase e que não altera seu significado pode ser mais adequadamente descrito como decorrente de uma transformação. A inserção de elementos redundantes deverá, pois, servir outros objetivos comunicativos-expressivos, que não a mera expressão da informação presente na forma da base. É o que sucede, por exemplo, com a topicalização ou a extração *ser...que* (clivagem), que correspondem a formas de tornar mais saliente um elemento da construção de base, ou mesmo a construção passiva, que degrada a saliência do agente e torna mais proeminente o objeto ou o paciente desse predicado.

As frases com o complemento de eco são efetivamente mais complexas do que a forma sem esse complemento, a despeito de carregarem a mesma informação. Quando a completiva-sujeito antecede o verbo copulativo, o complemento de eco deve, necessariamente, ser pronominalizado, como se vê pela aceitabilidade de (2a) e inaceitabilidade de (2b).

- (2) a. *Que o João_i faça isso é incomum (da parte dele_i + de sua_i parte)*
 b. * *Que o João_i faça isso é incomum da parte do João_i*

Quando a completiva-sujeito se encontra deslocada para o final da frase (*permuta de comprimento*, [Harris 1976]), o elemento reduzido a pronome deve ser um dos argumentos dessa oração subordinada, o que se exemplifica com (3a). A frase sem pronominalização, (3b), é inaceitável:

- (3) a. *É incomum da parte do João_i que ele_i faça isso*
 b. * *É incomum da parte do João_i que o João_i faça isso*

Para completivas-sujeito reduzidas de infinitivo (4a)-(4c), há ainda a possibilidade de redução a zero do sujeito da completiva, o que leva a formas como (4c).

- (4) a. *Que o João_i faça isso é interessante para ele_i*
 b. *= O João_i fazer isso é interessante para ele_i*
 c. *= Fazer isso_i é interessante para o João_i*

Assim, as frases de base dessas construções, como (1a), são representadas pela notação $(Que N^0 V ((Prep) N^1))_0 V_{cop} Adj (W)$, em que: N^0 representa o sujeito da completiva; V representa seu verbo principal; os elementos opcionais $Prep N^1$ representam o complemento da completiva; V_{cop} corresponde ao verbo copulativo; Adj se refere a um adjetivo predicativo; e W se refere a um complemento opcional do adjetivo. Tal sequência serve como forma de base das construções sobre as quais operam as inserções discutidas neste artigo.

As frases construídas que utilizamos como exemplo ao longo deste texto são baseadas em dados da partição brasileira do corpus PtTenTen20, acessado pela plataforma SketchEngine². Trata-se de um corpus de textos escritos de gêneros diversos extraídos da web no ano de 2020. A partição brasileira contém 8.010.603.604 *tokens*.

As seções seguintes deste artigo apresentam os três tipos de complemento de eco que analisamos como inserções nessa frase de base, respectivamente: *da parte de* N^0 (Seção 2), *para com* N^1 (Seção 3) e *para* N^0 (Seção 4). O artigo se conclui (Seção 5) com uma síntese das observações realizadas.

2. *Que* $N^0 V W V_{cop} Adj$ *da parte de* N^0

A primeira transformação aqui abordada consiste na equivalência entre uma frase com completiva-sujeito e a estrutura (A) *Que* $N^0 V W V_{cop} Adj$ *da parte de* N^0 , como exemplificado em (5a-5b), em que o sujeito da completiva-sujeito é correferente ao nome introduzido por *da parte de* (5c). Por essa razão, esse nome é normalmente pronominalizado ou o sintagma preposicional *de* N^0 é reduzido a um pronome possessivo (notado D_{poss}).

- (5) a. *Que o João faça isso é cruel*
 b. *Que o João_i faça isso é cruel (da parte de _ele_i + da sua_i parte)*
 c. **Que o João_i faça isso é cruel (da parte de _ela_j + da tua_k parte)*

Há uma variante menos frequente³, *por parte de*, à qual também se aplica a discussão a seguir:

² <https://www.sketchengine.eu/>, último acesso em 06/10/2024

³ Na partição brasileira do corpus PtTenTen2020 [Wagner Filho et al. 2018], observa-se 281 ocorrências da sequência *ser (Adv) Adj por parte de*, contra 1.390 da sequência *ser (Adv) Adj (da+de) (D_{poss}) parte*. As primeiras incluem frases como a seguinte (ênfase nosso): *Em uma guerra onde um continente está sendo dominado por uma nação armada, é pouco inteligente por parte de qualquer um manter suas economias e riquezas em qualquer banco que seja de um país pertencente a este continente*.

(6) *Que o João_i faça isso é cruel por parte de ele_i*

As ocorrências de (A) podem ser analisadas como derivadas transformacionalmente da frase de base com completiva-sujeito, consistindo na inserção de complemento de adjetivo *da parte de N⁰*, em que *N⁰* retoma o sujeito da oração completiva. Ao tentar inserir argumentos não correferentes, obtemos ora expressões inaceitáveis, como em (5c), ora construções de baixa aceitabilidade e em que o argumento introduzido por *da parte de* é interpretado como tendo alguma forma de controle sobre o sujeito da completiva, fato já notado por [Picabia 1978, p. 102]:

(7) ? *Que a empresa faça isso é cruel da parte do João*

Aceitando-se a frase (7), é necessário interpretá-la como sendo *João* o responsável pelas ações da *empresa*, numa relação de *anáfora indireta (metonímia)*. Trata-se, também assim, de um complemento de eco, ao mesmo título que os casos de *anáfora direta (João_i = ele_i)* que vimos anteriormente.

Para a inserção desse complemento de eco na frase, é necessário que o argumento duplicado seja um **agente** na completiva-sujeito⁴. Assim, o complemento de eco é aceitável, por exemplo, na combinação *usar/bizarro*:

- (8) a. *Que o João use essa roupa é bizarro*
 b. *Que o João_i use essa roupa é bizarro da parte dele_i*

Observamos frases de baixa aceitabilidade ao tentar inseri-lo quando a completiva-sujeito traz um verbo como *cabem*, que pede um sujeito **não agente**:

- (9) a. *Que o João caiba nessa roupa é bizarro*
 b. **Que o João_i caiba nessa roupa é bizarro da parte dele_i*

3. *Que N⁰ V (Prep) N¹ V_{cop} Adj para com N¹*

Nesta seção tratamos das frases com completiva-sujeito (10a) que podem receber o complemento de eco *para com N¹* (10b).

- (10) a. *Que o João tenha dito isso à Maria é justo*
 b. *Que o João tenha dito isso à Maria_i é justo para com ela_i*

⁴Conceitos semânticos, nomeadamente, os papéis semânticos, são representados em **negrito**. Para determinar verbos com sujeito **não agente** e assim construir estes exemplos, recorreremos ao *Dicionário Gramatical de Verbos do Português* [Baptista and Mamede 2020].

Em [Baptista 2005, pp. 153–154], considera-se que, no português europeu, *para* e *com* são variações da locução preposicional *para com*, o que também é aplicável ao português do Brasil. No entanto, em nossos testes, optamos por utilizar apenas a locução *para com* para evitar as ambiguidades que as outras duas formas isoladas podem gerar: *para* também pode ser interpretado como equivalente a *na opinião de*, além de introduzir outro tipo de complemento de eco (ver seção 4); enquanto *com* pode assumir um sentido **causativo** com certos adjetivos.

A discussão da bibliografia em torno do assunto sugere que há mais de uma origem possível para a ocorrência desses complementos. Considera-se em [Ranchhod 1985] que *para com N* são complementos essenciais. [Baptista 2005, pp. 152–153] considera que alguns predicados nominais selecionam dois argumentos e têm uma interpretação *elíptica* quando as sequências *para com N* são omitidas (exemplos (11a)-(11b) do português europeu retirados de [Baptista 2005, p. 152]).

- (11) a. *Que o Zé faça isso é de uma grande amabilidade (para com a Ana)*
 b. *Que o Zé faça isso é de uma grande crueldade (para com a Ana)*

Nesses casos, tratar-se-iam de complementos dos nomes predicativos, ora com um papel semântico de **beneficiário** (11a), ora com um papel de **vítima** (11b), consoante polaridade, positiva ou negativa, respetivamente, desse predicado.

Na ausência destes complementos, os predicados expressos por estes nomes predicativos implicam sempre um **beneficiário/vítima**, o que pode ser expresso por um indefinido *para com alguém*. Por outro lado, na presença desses complementos, verifica-se, ainda, a correferência obrigatória (e a correspondente redundância) com um complemento da completiva, o que torna inaceitável expressões como (12) em que *N* não é referente ao complemento da completiva.

- (12) *Que o Zé faça isso à Ana_i é de uma grande amabilidade/crueldade para com ela_i/*o Rui*

Em outros casos, não se poderia dizer que *para com N* funciona como complemento do nome predicativo. Tal sucede no exemplo (13) com o nome *arbitrariedade*, caso que [Baptista 2005, pp. 157–162] considerou como complemento de eco (exemplo retirado deste autor):

- (13) *Que o Zé tenha proibido isso à Ana_i foi de uma enorme arbitrariedade para com ela_i/*o Rui*

Efetivamente, o nome predicativo *arbitrariedade* não seleciona um complemento, neste caso, de **vítima**, pois pode tomar para a sua completiva-sujeito outros predicados que não envolvam um complemento correferente ao nome do complemento de eco (14).

- (14) *Que o Zé tenha decidido isso (*à Ana_i) foi de uma enorme arbitrariedade (para com ela; /*o Rui)*

Nesse caso, a ausência do complemento de eco não acarreta o valor elíptico que víamos em (11a)-(11b). A presença do complemento de eco, por seu turno, implica que o facto de *o Zé ter decidido isso* tenha tido algum efeito sobre o *N* do complemento de eco, mesmo que este elemento não esteja presente na completiva-sujeito (ver atrás, anáfora indireta). Assim, o complemento *para com N* deixa-se analisar em (11a)-(11b) e (12) ora como complemento dos nomes *amabilidade* e *crueldade*, ora como complemento do verbo *fazer*, ao passo que em (14) deverá ser considerado um complemento de eco, retomando, mesmo que indiretamente, o conteúdo da completiva.

Já [Vivès 1982], apresenta propostas de derivação das formas com sujeito humano a partir de uma frase de base (15a) por meio de reestruturação (15b) e posterior redução do nome *attitude* ‘atitude’, considerado como apropriado ao adjetivo *ferme* ‘firme’ (15c) (exemplos retirados de [Vivès 1982, p. 229]):

- (15) a. *L ’attitude de Lea avec Max est ferme*
‘A atitude da Lea para com o Max é firme’
- b. *Lea est ferme dans son attitude avec Max*
‘A Lea é firme em sua atitude para com o Max’
- c. *Lea est ferme avec Max*
‘A Lea é firme para com o Max’

[Baptista 2005, pp. 158–159] levantou uma crítica a essa proposta: quando um nome predicativo se reduz, seus complementos também se reduzem, conforme demonstra a inaceitabilidade de (16c) (exemplos de [Baptista 2005, pp. 158–159]).

- (16) a. *Os comentários do Zé à política do Governo foram de uma enorme prudência*
- b. *O Zé foi de uma enorme prudência nos seus comentários à política do governo*
- c. ** O Zé foi de uma enorme prudência à política do Governo*

Tal leva a considerar que o complemento *avec N* em (15c) não pode ser um complemento de *attitude* ‘atitude’ (ou de qualquer outro nome apropriado reduzido).

Foi proposta uma análise similar [Meydan 1995, pp. 174–176] partindo de uma frase com oração subjetiva, que consiste na seguinte sequência de transformações:

- (17) a. *Que Luc se comporte ainsi avec Léa est méprisant de sa part*
‘Que o Luc se comporte assim para com a Léa é desprezível de sua parte’
- b. *Luc est méprisant de se comporter ainsi avec Léa*
‘O Luc é desprezível em se comportar assim para com a Léa’

- c. *Luc est méprisant avec Léa*
 ‘O Luc é desprezível para com a Léa’

Assim, tanto [Vivès 1982] quanto [Meydan 1995] preocupam-se em explicar a presença dos complementos *avec N* (*para com N*) em frases com sujeito humano. Não se encontrou proposta de análise da inserção dos complementos *para com N* em frases com completiva-sujeito de adjetivo. Baseando-se na proposta de tratar como complemento de eco alguns complementos de nome predicativo com completiva-sujeito [Baptista 2005], estendemos essa explicação também aos adjetivos.

É possível que a correferência entre o complemento da completiva e o complemento de eco não esteja explícita, como sucede em (18a). Porém, é necessário que de alguma maneira o complemento *para com N* esteja implicado naquilo que a completiva-sujeito descreve; esse complemento corresponde a um *dativo alargado*, isto é, um complemento circunstancial que relaciona um predicado a alguém que sofre suas consequências, como se observa em (18b) [Ranchhod 1985], [Baptista 2005, p. 156].

- (18) a. *Que o João abra a porta é justo para com o Pedro.*
 b. *Que o João abra a porta (para + a) o Pedro_i é justo para com ele_i*

4. *Que N⁰ V W V_{cop} Adj para N⁰*

Na seção anterior, observamos que *para* pode introduzir um complemento de eco correferente ao complemento da completiva-sujeito, em alternância com *com* e *para com*. Há ainda um outro complemento de eco, também introduzido por *para*, mas que retoma o sujeito da completiva e *não* admite a variação *para com/para/com*.

- (19) a. *Que o João faça isso é comum*
 b. *Que o João_i faça isso é comum para ele_i*

Ora, uma frase superficialmente idêntica a (19b) mas sem correferência entre o sujeito da completiva e o complemento *para N*, como ilustrada abaixo em (20a), é aceitável. No entanto, em tal frase o complemento *para N* não tem o mesmo sentido que verifica em (19).

- (20) a. *Que o João_i faça isso é comum para ela_j*
 b. *Que o João_i faça isso é comum, na opinião dela_j*

Trata-se, em vez disso, de um circunstancial de **opinião** equivalente a um complemento *na opinião de N* (20b). Em tais casos, consideramos que *para N* não constitui um complemento de eco.

Os complementos de eco *para N⁰* já haviam sido observados, sem que se lhes tivesse sido dada essa designação, por [Casteleiro 1981, p. 297], que os considerou relacionados à operação de *elevação de objeto*. É fato que as frases com complemento de eco podem sofrer elevação de objeto. Partindo-se de uma frase como (19b), que renumeramos abaixo como (21a), e na presença do complemento de eco, tem-se, em primeiro lugar, a redução da completiva-sujeito a oração infinitiva (21b), seguida da redução a zero de *João*, a redução do sujeito da oração infinitiva (21c) na presença do complemento de eco (*para ele*); e, finalmente, a operação elevação de objeto *isso* (21d), para sujeito da construção adjetival (*comum*), mantendo-se, naturalmente, o verbo da completiva (*de fazer*) como complemento desse adjetivo:

- (21) a. *Que o João faça isso é comum para ele*
 b. *O João fazer isso é comum para ele*
 c. *Fazer isso é comum para ele*
 d. *Isso é comum de fazer para ele*

Ora, a operação de elevação do objeto não é necessária para a caracterização do complemento de eco *para N*, uma vez que ele existe nas estruturas sem elevação, como já demonstrado em (19). Em vez disso, ele deve ser caracterizado como um complemento de eco da oração com completiva-sujeito.

Por fim, apesar de tanto *da parte de N⁰* quanto *para N⁰* trazerem uma repetição do sujeito da completiva-sujeito, esses complementos se aplicam a adjetivos distintos, de modo que não podem ser considerados variantes:

- (22) a. *Que o João_i faça isso é impossível (?? da parte dele_i + para ele_i)*
 b. *Que o João_i faça isso é cruel (da parte dele_i + * para ele_i)*

5. Conclusão

Com este artigo, apresentamos e descrevemos certos complementos de adjetivo – complementos de eco – que são correferentes a argumentos de suas completivas-sujeito. Propomos que esses complementos possam ser analisados como transformações incrementais sobre uma frase de base com completiva-sujeito e sem complemento, uma vez que as frases assim ligadas carregam a mesma informação fundamental. Em seguida, apresentamos as propriedades formais desses três tipos de complementos de eco do português brasileiro: *da parte de N*, *para com N* e *para N*. Em estudos subsequentes, pretendemos apresentar uma caracterização extensional dos adjetivos que aceitam tais complementos de eco, listando-os segundo suas propriedades, e propor uma classificação baseada nessas e em outras propriedades formais.

Agradecimentos

Ryan Saldanha Martinez: Este estudo foi financiado em parte pela Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001. Jorge Baptista desenvolveu sua pesquisa no Instituto de Engenharia de Sistemas e Computadores, Investigação e Desenvolvimento em Lisboa, INESC-ID Lisboa – Human Language Technology Laboratory (INESC-ID Lisboa/HLT) e foi parcialmente financiado pelos fundos nacionais por meio da Fundação para a Ciência e a Tecnologia (FCT), projeto UIDB/50021/2020 (DOI:10.54499/UIDB/50021/2020). Oto Araújo Vale e Ryan Saldanha Martinez: Parte deste trabalho foi realizado no âmbito do Centro de Inteligência Artificial da USP (C4AI -<http://c4ai.inova.usp.br/>), que tem o apoio da IBM e da FAPESP (processo 2019/07665-4). Este projeto também foi apoiado pelo Ministério da Ciência, Tecnologia e Inovações, que tem recursos da Lei nº 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.

References

- Baptista, J. (2005). *Sintaxe dos predicados nominais com ser de*. Fundação Calouste Gulbenkian & Fundação para a Ciência e a Tecnologia, Lisboa.
- Baptista, J. and Mamede, N. (2013). Reciprocal Echo Complements in Portuguese: Linguistic Description in view of Rule-based Parsing. In Baptista, J. and Monteleone, M., editors, *Proceedings of the 32nd International Conference on Lexis and Grammar (CLG'2013)*, pages 33–40, Faro, Portugal. CLG'2103, Universidade do Algarve – FCHS.
- Baptista, J. M. E. and Mamede, N. (2020). *Dicionário gramatical de verbos do português*. Universidade do Algarve Editora, Faro.
- Casteleiro, J. M. (1981). *Sintaxe transformacional do adjetivo: regência das construções completivas*. INIC, Lisboa.
- Guillet, A. and Leclère, C. (1981). Restructuration du groupe nominal. *Langages*, 63:99–125.
- Harris, Z. (1976). *Notes du cours de syntaxe*. Éditions du Seuil, Paris.
- Harris, Z. S. (1970). *The elementary transformations*, pages 482–532. Springer.
- Harris, Z. S. (1982). *A Grammar of English on Mathematical Principles*. John Wiley and Sons, Inc., New York.
- Harris, Z. S. (1991). *Theory of Language and Information. A Mathematical Approach*. Clarendon Press, Oxford.
- Meydan, M. (1995). *Transformations des constructions verbales et adjectivales: élaboration du lexique-grammaire des adjectifs déverbaux*. PhD thesis, Paris 7.
- Picabia, L. (1978). *Les constructions adjectivales en français: systématique transformationnelle*, volume 11. Librairie Droz.
- Ranchhod, E. (1985). A romance construction with constrained coreference. *Linguisticae Investigationes*, 9(2):343–363.

Vivès, R. (1982). Une analyse possible de certains compléments prépositionnels. *Linguisticae Investigationes*, 6(1):227–233.

Wagner Filho, J. A., Wilkens, R., Idiart, M., and Villavicencio, A. (2018). The brWaC corpus: a new open resource for Brazilian Portuguese. In *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018)*.

Modelagem baseada em frames para identificação do léxico da Violência de Gênero

Lorena Tasca Larré¹, Tiago Timponi Torrent²

¹Universidade Federal de Juiz de Fora (UFJF)

²Universidade Federal de Juiz de Fora (UFJF)

lorena.tasca@estudante.ufjf.br, tiago.torrent@ufjf.br

Abstract. *Research indicates alarming data on gender-based violence in Brazil and around the world, and suggests that this violence is underreported. This work consists of an effort to model the domain of violence for frame-based semantic analysis, in relation to the open fields of SINAN, which contain information on situations of violence in medical care. Frame-based semantic analysis seeks to assist the task of computational algorithms, identifying what type of frames were evoked in the related domain. After modeling, the analysis of the prevalence of frames and lexical units in the e-SUS APS and SINAN corpora allows the identification of possible trajectories of violence against women.*

Resumo. *Pesquisas apontam dados alarmantes da Violência de Gênero no Brasil e no mundo, bem como sugerem que há subnotificação dessa Violência. Este trabalho consiste em um esforço de modelagem do domínio da violência para análise semântica baseada em frames, no que se refere aos campos abertos do SINAN, os quais contêm informações sobre situações de violência nos atendimentos médicos. A análise semântica baseada em frames procura auxiliar a tarefa de algoritmos computacionais, identificando que tipo de frames foram evocados no domínio relacionado. Após a modelagem, a análise de prevalência de frames e unidades lexicais nos corpora do e-SUS APS e do SINAN permite identificar possíveis trajetórias da violência contra a mulher.*

1. Introdução

O presente artigo objetiva descrever o esforço de modelagem dos frames do domínio da violência, assim como o resultado demonstrativo dos léxicos evocadores dos domínios da saúde e da violência, que são pesquisas vinculadas ao projeto “Data linkage e análise textual baseada em frames para identificação de casos de violência de gênero em territórios”. O projeto da FrameNet Brasil em parceria com a Vital Strategies obteve dados do sistema de saúde e-SUS APS, do SINAN (Sistema Nacional de Agravos de Notificação) e do SIM (Sistema de Informações sobre Mortalidade)¹ das cidades de São Caetano do Sul/SP e Recife/PE para *linkage* dos campos parametrizados, que normalmente contêm informações diversas, incluindo identificação e endereço das pacientes, além da análise semântica baseada em frames dos campos abertos, que, no e-

1 Os dados obtidos no SIM (Sistema de Informações sobre Mortalidade) são parametrizados e foram utilizados apenas pela Vital Strategies para *linkage* com dados dos outros sistemas, com a finalidade de obter informações sobre óbito relacionados às vítimas de violência (com registros no SINAN). Neste trabalho não foi feita análise de dados do SIM.

SUS, contêm as principais queixas relativas à ida de uma paciente ao atendimento em saúde e, no SINAN, relatam a situação de violência. Os dados dos campos presentes no e-SUS e no SINAN motivam a modelagem do domínio da violência e a análise dos léxicos evocadores dos domínios saúde e da violência que serão relatadas neste artigo. A modelagem do domínio da saúde foi realizada anteriormente e descrita em Dutra, 2023 (Dutra *et al.*, 2023). É importante mencionar as motivações do trabalho: a OMS² entende a Violência de Gênero como um problema de saúde pública, sendo que uma em cada três mulheres já sofreu violência física ou sexual. De acordo com Simister (2012), pelo menos um bilhão de mulheres já sofreu violência de gênero. No Brasil³, conforme divulgação feita pelo Governo Federal, somente no primeiro semestre de 2022, foram registradas mais de trinta e uma mil denúncias de violência doméstica. Diversas pesquisas (Garbin *et al.*, 2015; Kind *et al.* 2013) apontam que há subnotificação dos casos de violência de gênero, de forma que estudos que possivelmente ajudem a identificar esses casos já no atendimento em saúde se mostram importantes e necessários.

2. Contextualização da Violência de Gênero no Corpus e Motivações do Estudo

No Brasil, a notificação dos casos de violência de gênero é obrigatória. Entretanto, muitos pesquisadores (Garbin *et al.*, 2015; Kind *et al.*, 2013) apontam que há subnotificação de casos de GBV, devido ao fato de muitas mulheres não procurarem os serviços de saúde quando sofrem violência. Além do mais, mesmo quando há procura, há dificuldade por parte dos profissionais de saúde em realizar o registro, devido a vários fatores, tais como sobrecarga, medo de retaliação por parte dos agressores, falta de informações adequadas sobre a importância da denúncia, entre outros. Além do SINAN, o sistema público de saúde no Brasil conta, ainda, com outras bases de dados nas quais podem ser inseridas informações decorrentes de episódios de GBV, que são: o Sistema e-SUS Atenção Primária à Saúde e o Sistema de Informação de Mortalidade (SIM). Os dois últimos, foram analisados pela própria Vital Strategies e por Dutra (2023).

A hipótese que motiva a análise semântica dos dados encontrados nos prontuários médicos conjuntamente com os dados provenientes das notificações de agravo e de mortalidade é a de que a mulher que sofreu violência doméstica, possivelmente, terá procurado atendimento médico, que terá sido registrado em prontuários eletrônicos da atenção básica no e-SUS e que, posteriormente, caso tenha infortunadamente vindo a óbito, haverá, por conseguinte, um registro no SIM. Levando em conta que esses sistemas não são integrados pelo governo, o projeto “Data linkage e análise textual baseada em frames para a identificação de casos de Violência de Gênero em territórios” visa utilizar essa integração dos dados obtidos por esses sistemas, bem como análise textual baseada em frames para identificar padrões que demonstrem a frequência e a probabilidade de ocorrer violência de gênero em determinado território. Dessa maneira, a modelagem inicial do domínio da saúde (Dutra *et al.*, 2023), a posterior modelagem do domínio da violência, bem como os léxicos evocadores

²Disponível em: https://www.who.int/health-topics/violence-against-women#tab=tab_1

³ Disponível em: <https://brasil.un.org/pt-br/115652-oms-uma-em-cada-3-mulheres-em-todo-o-mundo-sofre-violencia>

presentes em cada domínio se mostram essenciais para o estudo.

3. FrameNet e FrameNet Brasil – Aplicação teórica

Baseada na Semântica de Frames, a FrameNet é um recurso lexicográfico que surgiu em 1997, no International Computer Science Institute (ICSI), em Berkeley, e que inicialmente aplicou a teoria da Semântica de Frames (Fillmore, 1982) para a análise das possibilidades lexicográficas de itens lexicais em inglês (Fillmore & Baker, 2009). Sua metodologia é guiada pela teoria da Semântica de Frames (Fillmore, 1982, 1985), e as análises, sustentadas por evidências em *corpora*. Já a FrameNet Brasil é o projeto responsável pela expansão da FrameNet para o português brasileiro.

Relacionada a este projeto, a teoria da Semântica de Frames é aplicada no que concerne à anotação baseada em Frames, feita nos campos abertos dos sistemas do e-SUS (Dutra *et al.*, 2023) e SINAN, supracitados.

Na modelagem do domínio da violência, consideramos, por exemplo, um item lexical, como “conflito”. Para entender adequadamente o sentido deste item, devemos considerar uma cena, ou frame, no qual dois lados estão em conflito por determinado assunto, com uma finalidade, que poderia ser um consenso ou um favorecimento à vontade de um lado específico. Dessa forma, os itens: *Assunto*, *Finalidade*, *Lados*, *Lado 1* e *Lado 2* se colocam como itens indispensáveis ao acontecimento da cena e são elementos de frame nucleares do frame Encontro_hostil, evocado pela Unidade Lexical (UL) *conflito.n*.

Além disso, os frames, ou cenas, evocados por Unidades Lexicais como a exemplificada acima são interligados a outros frames que correspondem a cenas semelhantes, formando uma rede de frames ou FrameNet.

Figura 1. Frame Encontro_hostil

Encontro_hostil @Lexical @Generic @Violence #86 Hostile_encounter [en]

Definition
Este frame consiste em palavras que descrevem um encontro hostil entre forças opostas **Lado_1** e **Lado_2**, conceptualizadas coletivamente como **Lados** sobre um Problema disputado e / ou para atingir uma **Finalidade** específica.

Frame Elements

Core

- Assunto**
CoreSet: Finalidade
Uma questão não resolvida sobre a qual os dois lados de um encontro hostil estão em desacordo.
- Finalidade**
O resultado desejado do resultado do encontro hostil para o **Lado_1** ou para todos os **Lados** coletivamente. Pode se referir diretamente ao estado de coisas desejado, ou o estado de coisas pode estar implícito em uma entidade (abstrata ou concreta) que eles desejam.
- Lados**
Os lados expressos em conjunto em um encontro hostil.
- Lado_1**
Excludes: Lados
Requires: Lado_2
Um dos dois participantes em um encontro hostil.
- Lado_2**
Excludes: Lados
Requires: Lado_1
Um dos dois participantes em um encontro hostil, geralmente o segundo mencionado.

4. Modelagem do Domínio da Violência

A modelagem de um domínio e criação de novos frames pode ser feita a partir de dois tipos de metodologias: a *top-down* e a *bottom-up*.

Na metodologia *bottom-up*, o anotador parte dos dados para criar o frame. A primeira tarefa é selecionar um agrupamento de termos que se relacionem semanticamente. Em um segundo momento, parte-se para a pesquisa dos itens lexicais selecionados nos *corpora*, tendo em mente um possível frame para investigação. Após este estudo, selecionam-se sentenças com esses itens lexicais com o objetivo de analisar seu respectivo comportamento sintático-semântico. Assim, procuram-se regularidades que permitam a estruturação de uma cena específica, essa etapa é a definição do frame. Nela, as Unidades Lexicais (ULs) são selecionadas pelo analista, bem como é especificada a nuclearidade dos Elementos de Frame (Efs), assim como a relação entre os Efs.

No método *top-down*, as primeiras análises são delineadas pela intuição de um falante nativo ou conhecedor da língua que, como anotador, propõe uma definição para o frame, elencando as possíveis Unidades Lexicais que o evocam, designando a nuclearidade de cada elemento, traçando, assim, as possíveis relações entre frames e Efs. Em um segundo momento, verifica-se como foi o seu comportamento nos *corpora* para confirmar se essa definição condiz com os dados. Se necessárias, alterações na definição do frame e dos Efs podem ser realizadas.

No domínio da violência, ambas as metodologias foram utilizadas. A metodologia *top-down* foi necessária quando modificamos o frame preexistente Violência para o frame não lexical Cenário_da_Violência, para que abrangesse melhor o domínio. No caso do domínio da Violência, vários frames nucleares relacionados ao assunto já existiam na base de dados da FrameNet Brasil, tais como: Abusar, Arma, Morte, Matar e Experienciar_Ferimento_Corporal, de forma que eles foram listados para que se agrupassem neles as Unidades Lexicais correspondentes. No que corresponde à metodologia *bottom-up*, foram criados cinco novos frames para o domínio: Estupro, Pessoas_por_estado_transitório, Pessoas_por_enquadramento_criminal, Estar_ferido, Acidente, Descrição_de_acidente e Tráfego. No geral, para o domínio da Violência, foram adicionadas ULs em trinta e oito frames já existentes, totalizando quarenta e três frames e 1768 ULs que integram o domínio. Os dados dos campos abertos utilizados para modelagem de frames passou por devida anonimização.

5. Fenômenos Capturados pelo Modelo

Para apurar como o léxico da Violência se manifesta nos campos abertos dos sistemas utilizados para registros, foi feita anotação automática por meio do LOME (Xia *et al.*, 2021), um sistema multilíngue que implementa um analisador semântico (*parser*), dentre outras possibilidades de processamento de língua natural. A partir das anotações geradas por esse *parser*, foi possível gerar um relatório de frames mais frequentes e ULs mais relevantes para os dados das vítimas de violência. De acordo com este relatório, apresentamos os léxicos evocadores de ambos os domínios modelados.

Para obtenção de uma lista de frames mais frequentes, a anotação automática via *parser* em *subcorpus* relacionado às vítimas de violência nos retornou alguns resultados. Assim, em comparação com o *corpus* geral de ambos os sistemas, foi obtido o índice *keyness* que corresponde a quantas vezes determinado frame ou UL é mais frequente naquele *corpus* do que no *corpus* geral. Dessa forma, no léxico evocador do domínio da saúde para o *subcorpus* e-SUS das vítimas de violência, temos o frame Intoxicantes como mais evocado, cuja UL mais evocadora é *chumbinho.n*, levando em

conta que este poderia ser ingerido pelas vítimas ou utilizado como veneno contra elas. Na tabela 1 temos alguns dos frames mais evocados com respectivos *keyness* e ULs mais frequentes para o domínio da saúde em *corpus* e-SUS:

Tabela 1. Frames e ULs mais frequentes do domínio da saúde no e-SUS

Frame	Unidades Lexicais	Keyness
Intoxicantes	<i>Chumbinho.n, entorpecente.n, crack.n</i>	2,9
Vício	<i>Viciado.a, viciar.v, drogado.a</i>	2,55
Substância_tóxica	<i>Veneno de rato.n, veneno de barata.n, veneno.n</i>	1,63
Patógenos	<i>Tuberculose.n, vírus.n, ameba.n</i>	1,55
Institucionalização	<i>Admissão.n, admitir.v, internar.v</i>	1,36

Temos, então, para o domínio da saúde no *corpus* e-sus de vítimas da violência, em comparação com *corpus* genérico, uma alta frequência de frames e Unidades Lexicais relacionadas a abuso de substâncias tóxicas ou tentativa de envenenamento, sendo via suicídio ou ato criminoso. Na tabela 2, veremos frames menos frequentes (com menor *keyness*), mas que demonstram situações específicas possivelmente relacionadas à violência de gênero.

Tabela 2. Frames e ULs relevantes no domínio da saúde no e-SUS

Frame	Unidades Lexicais	Keyness
Medicamentos	<i>Prebictal.n, hadol.n, cartela.n</i>	1,04
Condições_em_saúde	<i>Intoxicação exógena.n, intoxicação medicamentosa.n, envenenamento.n</i>	1,04
Sintomas	<i>Apatia.n, défict auditivo.n, boca seca.n, úlcera.n, pensamentos suicidas.n</i>	0,99
Descrição_de_sintoma_ou_condição_em_saúde	<i>Psicológico.a</i>	0,96
Especialidades_em_saúde	<i>Psiquiátrico.a</i>	0,88

Em síntese, os frames mais evocados e ULs mais evocadoras do domínio da saúde no *corpus* de vítimas de violência do e-SUS sugerem uma história de uma paciente em situação de ansiedade, que pode apresentar transtornos psiquiátricos possivelmente decorrentes de atos de violência prévios, de discussões e outras situações de descontentamento.

Se tratando do léxico evocador do domínio da violência, ainda em *corpus* e-SUS, temos o frame *Disparar_projéteis* com maior índice *keyness*, ou seja, este frame aparece 6,05 vezes mais no *subcorpus* relativo aos atendimentos realizados com vítimas

de violência contra a mulher. A UL mais evocadora desse frame, sendo que ela aparece 21,7 vezes mais do que no corpus geral, é o verbo *atirar.v*. Na tabela 3, podemos ver alguns dos frames e ULs mais frequentes para o domínio da violência no *corpus* e-SUS.

Tabela 3. Frames e ULs mais frequentes do domínio da da violência, no e-SUS

Frame	Unidades Lexicais	Keyness
Disparar_projéteis	<i>Ativar.v, tiroteio.n</i>	6,05
Abusar	<i>Assediar.v, alienaçãoparental.n, abuso.n</i>	5,92
Cenário_da_violência	<i>Violência.n</i>	5,9
Matar	<i>Homicidar.v, homicídio.n, suicídio.n</i>	5,37
Pessoa_por_enquadramento_criminal	<i>Traficante.n</i>	5,14

Temos, então, para o domínio da violência no *corpus* e-SUS de vítimas, em comparação com *corpus* genérico, uma alta frequência de frames e Unidades Lexicais relacionadas a disparo de arma de fogo, abuso e assassinato. Na tabela 4, veremos frames menos frequentes (com menor *keyness*), mas que também demonstram situações específicas possivelmente relacionadas aos diversos aspectos de como a violência ocorre.

Tabela 4. Frames e ULs relevantes do domínio da violência, no e-SUS

Frame	Unidades Lexicais	Keyness
Estupro	<i>Estupro.n, estuprar.v</i>	2,41
Encontro_hostil	<i>Batalha.n, palavra.n, enfrentamento.n</i>	2,05
Causar_dano	<i>Apedrejar.v, injuriar.v, socar.v</i>	1,66
Decisão_judicial	<i>Medida protetiva.n</i>	1,65
Arma	<i>Navalha.n, arma branca.n, faca.n</i>	1,25

Como podemos perceber pelos frames mais evocados e pelas ULs que mais evocam com relação ao domínio da violência no *corpus* e-SUS, a violência contra mulher, muitas vezes, ocorre de maneira extrema, com disparo de projéteis, outras vezes por meio de violência física ou de diversos tipos de armas brancas, sendo que também pode ocorrer na forma de abuso, ameaça, exploração sexual e estupro, por exemplo, podendo ter como fatores predeterminantes alguns dos itens já relatados anteriormente, como ansiedade, transtornos psiquiátricos e abuso de entorpecentes por ambas as partes.

Por fim, temos o léxico evocador dos domínios da saúde e da violência nos campos abertos do SINAN, demonstrando como a violência de gênero chega aos

registros desse sistema. Para o domínio da saúde temos como frame mais frequente o Causar_condição_em_saúde, apresentando *keyness* 36,35, ou seja, sendo 36 vezes mais frequente do que em *corpus* genérico. A UL mais frequente desse frame é *ingerir.v*, corroborando a ideia de tentativa de suicídio ou de envenenamento das vítimas. Na tabela 5, vemos os cinco frames mais frequentes e suas respectivas ULs.

Tabela 5. Frames e ULs mais frequentes do domínio da saúde, no SINAN

Frame	Unidades Lexicais	Keyness
Causar_condição_em_saúde	<i>Ingerir.v</i>	36,3
Vício	<i>Drogado.n, drogado.a, viciado.n</i>	29,2
Substância_tóxica	<i>Veneno de rato.n, veneno de barata.n, veneno.n</i>	21,5
Intoxicantes	<i>Chumbinho.n, substância psicoativa.n, bala.n</i>	15,6
Institucionalização	<i>Ingerido.a, admitir.v</i>	13,5

Para elucidarmos outros fenômenos capturados com relação ao domínio da saúde em *corpus* SINAN, é importante citar outros frames também muito evocados nesse *corpus*, como o frame Pessoas_por_condição_em_saúde, com índice *keyness* de 3,9. Nesse frame, não só a UL *paciente.n* é evocadora, assim como a UL *vítima.n*. Os demais frames, apesar de não serem os mais evocados, demonstram particularidades relativas às situações de violência e podem ser verificados na tabela 6.

Tabela 6. Frames e ULs relevantes no domínio da saúde, no SINAN

Frame	Unidades Lexicais	Keyness
Pessoas_por_condição_em_saúde	<i>Paciente.n, vítima.n</i>	3,9
Condições_em_saúde	<i>Alcoolizado.a, corte.n, envenenamento.n</i>	1,2
Medicamentos	<i>Cartela.n, comprimidos.n, rivotril.n</i>	1,1
Partes_do_corpo	<i>Cara.n, hemiface.n, rosto.n</i>	0,83
Descrição_orientada_pela_parte_do_corpo	<i>Periorbital.a, craniano.a</i>	0,6

Na tabela 6, podemos verificar também como o abuso de intoxicantes está relacionado aos atos de violência, assim como podemos perceber a prevalência de Unidades Lexicais relativas ao uso de medicamentos envolvendo envenenamento ou suicídio, como podemos exemplificar pelas sentenças adaptadas extraídas do *corpus*: “Encontrada desacordada ao lado de cartela de remédios”; ou “Sofreu tentativa de

envenenamento por rixa familiar”. Nos frames Partes_do_corpo e Descrição_orientada_pela_parte_do_corpo, temos ULs como *cara.n*, *hemiface.n* e *craniano.a* como mais frequentes, demonstrando a tendência de as agressões ocorrerem na região da cabeça.

Relativo ao domínio da violência em *corpus* SINAN, observando a comparação com *corpus* genérico, temos o frame Pessoas_por_enquadramento_criminal como mais frequente, apresentando índice *keyness* 111,8 ou seja, sendo evocado quase 112 vezes mais. Outros frames mais frequentes podem ser vistos na tabela 7.

Tabela 7. Frames e ULs mais frequentes do domínio da violência, no SINAN

Frame	Unidades Lexicais	Keyness
Pessoas_por_enquadramento_criminal	<i>Autor.n, agressor.n, assaltante.n, acusado.n</i>	111,8
Disparar_projéteis	<i>Atirar.v, tiroteio.n</i>	54,9
Cenário_da_violência	<i>Violência.n</i>	39,43
Matar	<i>Homicidiar.v, suicidar.v, matar.v</i>	13,08
Estupro	<i>Estupro.n, estuprar.v</i>	9,59

Os frames mais evocados para o domínio da saúde no SINAN, conforme podemos ver na tabela 7 já demonstram as cenas com o agressor e com o tipo de violência, que pode envolver arma de fogo, assassinato e estupro, por exemplo. Outro frame relativamente frequente para esse domínio e que também se mostra importante é o frame Abusar, que é 8,64 vezes mais evocado. Para esse frame é importante notar que a UL mais evocadora é *alienação parental.n*, evocando 39,72 vezes mais, seguida das ULs *ameaçado.a*, *assediar.v*, *abusar.v*, *abuso.n*, *perseguir.v*, *exploração sexual.n* e *abusivo.a*, que são de 5,27 a 27,26 vezes mais evocadoras. Na tabela 8, vemos este e outros frames relevantes.

Tabela 8. Frames e ULs relevantes do domínio da violência, no SINAN

Frame	Unidades Lexicais	Keyness
Abusar	<i>Alienação parental.n, ameaçado.a, assediar.v</i>	8,64
Arma	<i>Porrete.n, arma branca.n, pau.n</i>	6,94
Operar_veículo	<i>Pilotar.v</i>	6,5
Causar_dano	<i>Abusar.v, acertar.v, cotovelar.v</i>	4,93
Experienciar_ferimento_corporal	<i>Coronhada.n, espancamento.n, agredido.a</i>	4,83

Para a tabela 8, podemos ver relacionado ao domínio da violência no *corpus* SINAN, mais frames que denotam a concretização da violência, seja pelo alto índice do frame Abusar com a UL *Alienação parental.n*, demonstrando problemas conjugais e possível

violência sofrida por meninas (menores), assim como a UL *abusar.v* é retomada no frame *Causar_dano*, o que indica o contexto variado da ocorrência de abuso. Temos ainda a reincidência do frame *Arma*, indicando também o alto índice de uso de armas brancas e o frame *Operar_veículo*, que é evocado em situações de atropelamento (conforme consulta no *corpus*). Por último, o frame *Experienciar_ferimento_corporal* indica também outras formas de como a violência pode ocorrer.

6. Conclusões

O projeto se propôs ser um trabalho que, para além de descrever dois domínios lexicais do português brasileiro a partir da metodologia da FrameNet Brasil, pudesse fazer dessa descrição uma contribuição para um produto tecnológico que busca ajudar no enfrentamento de um problema social evidente, que é a violência contra as mulheres.

O estudo dos campos abertos tendo como base a semântica de frames facilita a contagem, o agrupamento e processamento dos dados, possibilitando um olhar mais abrangente sobre frames e unidades lexicais mais frequentes. Espera-se que o esforço de modelagem apresentado neste trabalho contribua sendo uma nova análise dos *corpora* trabalhados aqui e que as análises feitas e demonstradas possam ser ainda mais estudadas, podendo contribuir para criação de novos mecanismos que ajudem na identificação precoce de casos de Violência de Gênero, ou que motivem outros tipos de trabalhos que envolvam gestores e entidades públicas na mitigação do problema da subnotificação e da Violência de Gênero como um todo.

7. Agradecimentos

A pesquisa desenvolvida por Lorena Larré foi financiada com bolsa CAPES/PROEX nº 88887.716453/2022-00. Tiago Torrent é bolsista de produtividade em pesquisa do CNPq, processo nº 313103/2021-6.

8. Referências

DUTRA, L.; LORENZI, A.; LARRÉ, L.; BELCAVELLO, F.; MATOS, E.; PESTANA, A.; BROWN, K.; GONÇALVES, M.; HERBST, V.; REINACH, S.; TEIXEIRA, R.; PAULA, P.; PELLINI, A.; SEQUEIRA, C.; SABINO, E.; LEAL, F.; CONDE, M.; GRESPAN, R.; TORRENT, T. (2023). Building a Frame-Semantic Model of the Healthcare Domain: Towards the identification of gender-based violence in public health data. *In: Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana (STIL)*, 14., 2023, Belo Horizonte/MG. Anais [...]. Porto Alegre: Sociedade Brasileira de Computação, 2023. p.338-346.

FILLMORE, C. J. (1982). Frame semantics. In: Linguistic Society of Korea (Ed.), "Linguistics in The Morning Calm". Seoul: Hanshin, p.111-138.

FILLMORE, C. J. (1985) Frames and the semantics of understanding. *Quaderni di Semantica*, 6(2), 222-254.

FILLMORE, C. J., BAKER, C. (2009). A frames approach to semantic analysis. In Heine, B. & Narrog, H. (Eds). *The Oxford Handbook of Linguistic Analysis*. Oxford: Oxford University Press.

GARBIN, C. A. S.; DIAS, I. A.; ROVIDA, T. A. S.; GARBIN, A. J. I. (2015). Desafios do profissional de saúde na notificação da violência: obrigatoriedade, efetivação e encaminhamento. *Ciência & Saúde Coletiva*, 20, 1879-1890.

KIND, L.; ORSINI, M. L. P.; NEPOMUCENO, V.; GONÇALVES, L.; SOUZA, G. A.; FERREIRA, M. F. F. (2013). Subnotificação e (in) visibilidade da violência contra mulheres na atenção primária à saúde. *Cadernos de Saúde Pública*, 29, 1805-1815.

SIMISTER, J. (2012). *Gender Based Violence - Causes and Remedies*. New York: Nova Science Publishers Inc.

Relações Retóricas de *List* e *Sequence* em textos jornalísticos

Tobias J. A. Almeida¹, Patrícia V. Almeida¹, Paula C.F. Cardoso²

¹FAELCH/ Departamento de Estudos da Linguagem – Universidade Federal de Lavras

²Faculdade de Computação – Universidade Federal do Pará (UFPA)

{tobias.almeida@estudante.ufla.br, patriciaalmeida@ufla.br,
pcardoso@ufpa.br}

Abstract. *This study investigates the rhetorical relations of List and Sequence in journalistic texts to identify discourse signals (DS) beyond conventional discourse markers (DMs). Utilizing a corpus of 87 texts from CSTNews, the research considers various elements, including morphological and syntactic aspects. Results reveal that other discourse signals significantly contribute to identifying these relations. The study concludes that integrating diverse discourse signals enhances discourse analysis and suggests new research directions.*

Resumo. *Esta pesquisa investiga as relações retóricas List e Sequence em textos jornalísticos, visando identificar sinalizadores discursivos (SDs) além dos marcadores discursivos (MDs) convencionais. Utilizando um corpus de 87 textos do CSTNews, a análise abrange elementos morfológicos, semânticos, sintáticos e gráficos. Os resultados mostram que outros SDs são essenciais para a identificação dessas relações. Conclui-se que a combinação de diferentes SDs enriquece a análise discursiva e sugere novas direções de pesquisa.*

1. Introdução

A *Rhetorical Structure Theory* (RST), ou Teoria da Estrutura Retórica em português, é uma teoria linguístico-descritiva que tem como foco a organização textual, caracterizando as relações estabelecidas entre as partes de um texto para formar uma estrutura discursiva coesa e interligada, frequentemente representada em forma de árvore (Mann e Thompson, 1987). As relações retóricas descritas pela RST indicam como essas partes se conectam para compor um todo coerente e significativo. A teoria concentra-se nas interações entre as unidades de informação que constituem um texto, evidenciando como sua organização contribui tanto para a expressão da intenção comunicativa do autor quanto para orientar a interpretação do leitor.

Comumente, essas relações são identificadas por meio de marcadores discursivos (MDs), que são elementos coesivos compostos por uma ou mais palavras que tornam explícita a relação entre as partes de um texto (Koch e Travaglia, 2002). No entanto, a presença de MDs nem sempre é clara ou suficiente para definir a relação retórica existente. Nesse contexto, Taboada e Das (2013; 2018) sugerem que a análise deve ir além dos MDs, incorporando outros elementos formais conhecidos como

sinalizadores discursivos (SD) e discutem a necessidade de identificar outros elementos, uma vez que nem todas as relações retóricas apresentam um MD de forma explícita.

Diante do exposto, este trabalho investiga as relações retóricas *List* e *Sequence*, que, de acordo com Mann e Thompson (1987), são multinucleares, ou seja, compostas por múltiplos núcleos de igual importância. Na relação *List*, os elementos apresentados são comparáveis e conectados por uma estrutura de lista, em que o leitor reconhece a possibilidade de comparação entre os elementos apresentados. Por outro lado, a relação *Sequence*, apresenta uma sucessão temporal de eventos, descritos de forma sequencial em que o leitor percebe na construção das proposições a sequência em que os fatos apresentados ocorreram. Nesse sentido, com base na noção de SDs e no estudo de Rodrigues, Souza e Cardoso (2023), o objetivo desta pesquisa é examinar quais SDs, restritos às proposições intrasentenciais, indicam a presença das relações *List* e *Sequence* em textos jornalísticos do português brasileiro (PB). Na seção de resultados, são apresentados exemplos de alguns dos SDs observados até o momento para essas relações.

2. Trabalhos relacionados

Nesta seção, destacam-se três trabalhos relacionados aos objetivos desta pesquisa.

Antônio e Rúbio (2012) analisaram um corpus de elocuições formais em PB, concluindo que as relações entre orações podem ser descritas pela intenção comunicativa do autor e sua avaliação do destinatário. Eles observaram que, na relação *List*, o conectivo “e” é usado para estabelecer a relação, sem subsequência temporal, e os elementos são comparáveis. Na relação *Sequence*, observaram o uso do conectivo "e" com conjugação verbal, indicando sequência temporal, além de conectivos de sequenciadores temporais.

Antônio (2017), usando um corpus oral, argumentou que as relações RST são baseadas em sentido, não apenas na forma, e são interpretadas independentemente de serem marcadas explicitamente por conectivos. Ele identificou diversos SD, incluindo pontuação, correlação modo-temporal, conteúdo textual, paralelismo sintático, paráfrase, inserção parentética, repetição e apresentação de evidências.

Para a língua inglesa, Das e Taboada (2018) analisaram os SD no corpus *RST Discourse Treebank*. Eles identificaram que uma relação pode ser marcada por um sinalizador único (como MD, referências pessoais, orações relativas) ou por um sinalizador combinado (como vírgula + oração no particípio passado, construção sintática paralela + cadeia lexical).

3. Metodologia

Para realizar esta pesquisa, foi utilizado um corpus de textos jornalísticos em PB, anotado manualmente em diversos aspectos da organização do discurso, como a RST. O corpus utilizado foi o CSTNews¹ (Cardoso et al., 2011), composto por 50 conjuntos de textos jornalísticos coletados em 2007, que abrangem diversos tópicos de agências de notícias online populares no Brasil.

Com foco nas relações discursivas de *List* e *Sequence*, foram selecionados 31 *clusters*, compostos por 2 ou 3 textos cada, resultando em um total de 87 textos. O

¹Disponível em: <http://nilc.icmc.usp.br/CSTNews/login/?next=/CSTNews/>

anotador, a partir da leitura e da investigação dos elementos presentes nos textos, identificou e anotou os SDs encontrados nas proposições. Baseando-se em estudos prévios, supramencionados, a tarefa consistia em identificar outros possíveis sinalizadores que ocorressem nas relações analisadas, considerando elementos morfológicos, sintáticos, semânticos e gráficos. Para isso, foi empregada a taxonomia de SDs catalogada no estudo de Dantas et al. (2024), com o objetivo de identificar as classificações existentes ou propor novas categorias a partir dessa taxonomia. A ferramenta rstWeb (Zeldes, 2016) foi utilizada para visualizar os textos em formato de árvore, facilitando a anotação dos SDs. Concluída essa etapa, foi realizada uma reflexão e análise dos elementos formais que constituem os sinalizadores discursivos, a fim de atingir os objetivos da investigação.

4. Discussão e resultados parciais

Nesta seção, são apresentados os resultados parciais obtidos a partir da análise geral dos dados coletados, com ênfase nas relações *List* e *Sequence*. A análise dos textos anotados considerou, além dos MDs, outros elementos formais como SDs, incluindo aspectos morfológicos, semânticos, sintáticos e gráficos, por serem pistas linguísticas e estruturais com grande potencial para indicar as relações propostas pela RST.

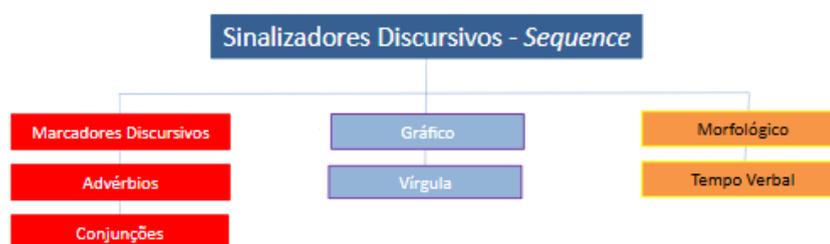
Na Figura 1, é apresentada uma proposta de categorização dos elementos encontrados para a relação de *List*.

Figura 1. Sinalizadores Discursivos de List



Na Figura 2, é apresentada uma proposta de categorização dos elementos encontrados para a relação de *List*.

Figura 2. Sinalizadores Discursivos de Sequence



Na Figura 3, são apresentados alguns exemplos, selecionados aleatoriamente dentro do conjunto de 87 textos, de SDs encontrados nos textos analisados.

Figura 3. Exemplo de *List* e *Sequence*

Relação	Sinalizadores	Exemplos
List	MD (advérbio) + gráfico (vírgula) + morfológico (sentido verbal) + paralelismo sintático + MD (conjunção) sintático (numeral)	(1) Neste cenário, Lula sobe de 27% para 31%, e Alckmin passa de 4% para 14%, Heloísa também cresce, de 1% para 6%.
	sintático (numeral) + morfológico (sentido verbal) + MD (conjunção)	(2) ...57 pessoas morreram e 128 ficaram feridas
Sequence	gráfico (vírgula) + morfológico (tempo verbal) + MD (advérbio) +	(5) Na política partidária, iniciou na extinta UDN, tendo sido eleito para um mandato de deputado estadual e dois de federal até 1967, até ser nomeado prefeito de Salvador.
	MD (advérbio) + MD (conjunção) + morfológico (tempo verbal)	(7) ... chegou na semana passada à ilha e manteve diferentes reuniões com as duas partes.

Os resultados apresentados na Figura 3, foram selecionados aleatoriamente a partir do com foco nas construções que continham o maior número de SDs. A seleção aleatória foi realizada com o intuito de exemplificar uma amostra representativa das construções que mais se destacavam pela presença de SDs. Tais resultados revelam que para ambas relações analisadas os SDs se constituem na combinação de MDs prototípicos já identificados pela literatura e outros elementos formais supramencionados. Embora a conjunção "e" seja frequentemente utilizada como um MD para ambas as relações, outros sinalizadores ajudam a distingui-las. Nas relações de *List*, observa-se uma igualdade entre os itens apresentados, que nem sempre seguem uma ordem temporal, sendo geralmente construídos em paralelismo sintático. Já nas relações de *Sequence*, a sucessão temporal dos eventos é enfatizada, frequentemente indicada por sinalizadores morfológicos, como o tempo verbal, e por MDs, como advérbios temporais, que detalham a ordem cronológica dos fatos.

Além disso, foi observado que a relação de *List*, nos exemplos analisados neste estudo (embora nem todos foram exemplificados aqui), é simétrica, ou seja, os elementos das orações podem ser trocados sem alterar o sentido. Em contraste, a relação de *Sequence* é assimétrica, pois indica uma sucessão temporal, o que impede a troca dos elementos sem modificar o sentido.

5. Considerações Finais

Neste estudo preliminar, o objetivo foi explorar pistas adicionais que pudessem indicar as relações de *List* e *Sequence* no modelo teórico da RST, utilizando um corpus previamente anotado. Os resultados ressaltam a importância de considerar sinalizadores discursivos além dos MDs, extrapolando a análise superficial a fim de ir mais a fundo na compreensão das relações analisadas.

Para pesquisas futuras, planeja-se ampliar a amostra e expandir as análises para incluir relações intersentenciais. Além disso, será relevante investigar outros gêneros textuais, como textos não estruturados (por exemplo, tweets), dado que o corpus utilizado nesta pesquisa é composto exclusivamente por textos jornalísticos.

Agradecimentos

Este trabalho foi realizado no âmbito do Centro de Inteligência Artificial da Universidade de São Paulo (C4AI -<http://c4ai.inova.usp.br/>), com o apoio da Fundação de Amparo à Pesquisa do Estado de São Paulo (processo FAPESP #2019/07665-4) e da IBM. Este projeto também foi apoiado pelo Ministério da Ciência, Tecnologia e Inovações, com recursos 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.

6. Referências

- Antonio, Juliano Desiderato. **Mecanismos utilizados pelos destinatários do discurso para identificação de relações de coerência não sinalizadas por conectores**. DELTA: Documentação e Estudos em Linguística Teórica e Aplicada, v. 33, n. 1, 2017.
- Antonio, Juliano Desiderato; RÚBIO, Marília Gabriela. **Relações retóricas de lista e de sequência estabelecidas por orações paratáticas aditivas e por orações paratáticas justapostas: uma investigação funcionalista**. Signum: Estudos da Linguagem, v. 15, n. 1, p. 33-56, 2012
- Das, D. e Taboada, M. (2018) **RST Signalling Corpus: A corpus of signals of coherence relations**. Language Resources and Evaluation, Vol 52, N. 1, pp. 149-184
- DANTAS, Ewerson et al. **Manual de anotação de sinalizadores discursivos em textos jornalísticos**. São Carlos: Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo. Disponível em: <https://repositorio.usp.br/item/003207370>. Acesso em: 25 ago. 2024. , 2024
- Koch, I.V. e TRAVAGLIA, L.C. (2002). **A Coerência Textual**. Editora Contexto.
- Mann, W. C.; Thompson, S. A. **Rhetorical structure theory: Toward a functional theory of text organization**. Text-interdisciplinary Journal for the Study of Discourse, v. 8, n. 3, p. 243–281, 1988.
- Rodrigues, R.; Souza, J.W.C; Cardoso, P.C.F.. **Sinalizadores retórico-discursivos: revisitando a anotação RST no cópurs CSTNews**. In: Anais do XIV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana. SBC, 2023. p. 249-257
- Souza, J.W.C.; Cardoso, P.C.F.; Rodrigues, R. **Systematic Review of Studies on Rhetorical Structure Theory (RST)**. Revista de Estudos da Linguagem, [s.l.], v. 31, n. 3, p. 1643-1675, 2023.
- Taboada, Maite; DAS, Debopam. **Annotation upon annotation: Adding signalling information to a corpus of discourse relations**. Dialogue & Discourse, v. 4, n. 2, p. 249-281, 2013.

Estudo preliminar sobre sinalizadores discursivos para Conteúdo Gerado por Usuários

Naira Silva Gama¹, Jackson Wilke da Cruz Souza^{1,2}

¹Instituto de Ciência, Tecnologia e Inovação
Universidade Federal da Bahia (UFBA), Camaçari/BA

²Programa de Pós-Graduação em Língua e Cultura (PPGLinC)
Universidade Federal da Bahia (UFBA)– Salvador/BA

nairagama@ufba.br, jackcruzsouz@gmail.com

Abstract. *Rhetorical Structure Theory (RST) is a theory that points out the rhetorical structure present in the text. Descriptive RST works are mostly dedicated to formal textual genres, resulting in a scarcity of works that explore the theory in User-Generated Content (UGC) texts. Therefore, the objective of this work is to investigate discursive signals (SDs) of RST relationships in UGC texts, more specifically in tweets from the financial market. To this end, 180 tweets were randomly selected from the DANTE-stocks corpus, which were analyzed manually, identifying the flags of the RST relationships previously noted. As a result, the typology of flags for Portuguese was updated with SDs specific to UGC texts.*

Resumo. *A Rhetorical Structure Theory (RST) é uma teoria que aponta a estrutura retórica presente no texto. Os trabalhos descritivos de RST majoritariamente se dedicam a gêneros textuais formais, resultando na escassez de trabalhos que explorem a teoria em textos de Conteúdo Gerado por Usuário (CGU). Assim, objetiva-se neste trabalho investigar sinalizadores discursivos (SDs) das relações RST em textos de CGU, mais especificamente em tweets do mercado financeiro. Para tanto, foram selecionados aleatoriamente 180 tweets do corpus DANTE-stocks, que foram analisados manualmente identificando os sinalizadores das relações RST anotadas previamente. Como resultado, atualizou-se a tipologia de sinalizadores para o Português com SDs específicos de textos de CGU.*

1. Introdução

A *Rhetorical Structure Theory* (RST) [Mann e Thompson 1988] é uma teoria que possibilita a visualização da estrutura retórica de um texto a partir de um conjunto de relações semânticas, e, com efeito, facilita análises textuais. [Souza, Cardoso e Rodrigues 2023] destacam que esse modelo teórico vem sendo utilizado em estudos linguísticos, diálogo e multimídia e análise de discurso, argumentação e escrita, além de aplicações de Processamento de Linguagem Natural (PLN), como avaliação de textos, sumarização e tradução automática.

Majoritariamente, as relações semânticas do modelo RST vem sendo identificadas com base em marcadores discursivos. Porém, em estudos recentes [Souza, Cardoso e Rodrigues 2023; Dantas *et al.* 2024] realizados no *corpus* CSTNews [Cardoso *et al.* 2011], chegou-se à proposição de uma série de sinalizadores linguísticos e estruturais que servem como potenciais indicadores de relações RST. Os resultados corroboram investigações internacionais na mesma temática [Taboada e Das 2013; Das e Taboada 2018; Liu e Zeldes 2022], além de indicarem particularidades que ocorreram apenas em língua portuguesa.

Ao final do estudo, Dantas *et al.* (2024) propuseram uma taxonomia em que os sinalizadores estão organizados em tipos e subtipos, a saber: (i) *Marcadores discursivos*, tendo como subtipos Preposição, conjunção e advérbio; (ii) *Morfológico*, tendo como subtipos Tempo verbal, Pronome relativo e Numeral; (iii) *Sintático*, tendo Orações relativas, Orações circunstanciais e Valência verbais como subtipos; (iv) *Semântico*, tendo como subtipos as relações Hiperonímia/Hiponímia, Sinonímia e Antonímia, Campo semântico, Conhecimento de mundo, Fonte (da informação), Verbo de comunicação, Sentido do verbo e Acrônimos; e (v) *Gráfico*, com os subtipos Travessão, Parênteses, Pontuação e Aspas.

Por terem analisado um *corpus* de texto jornalístico, o registro linguístico sob observação é o formal. Trabalhos da literatura recente indicam que diante de outros registros e modalidades da língua podem gerar outros sinalizadores discursivos [Antonio 2017; Fachada 2019; Pecuch 2021]. Assim, neste trabalho foi conduzido um estudo preliminar para encontrar sinalizadores específicos de textos de CGU. Para tanto, partiu-se de dois grandes objetivos: (i) evidenciar a amplitude em textos de CGU dos sinalizadores discursivos identificados por [Dantas *et al.* 2024]; (ii) destacar sinalizadores em potencial desse novo gênero textual sob observação.

Para tanto, este texto foi organizado em 4 seções, além desta Introdução. Na Seção 2, apresentam-se a metodologia e o *corpus* utilizado a partir dos principais objetivos deste trabalho. Na Seção 3, apresentam-se os resultados, com uma análise preliminar do conjunto de dados, os sinalizadores identificados, a uma breve comparação entre a tipologia proposta e as de outros trabalhos. Por fim, na Seção 4, as considerações finais são expostas juntamente com os trabalhos futuros.

2. Metodologia

Neste trabalho, foi utilizado o *corpus* DANTE-stocks [Di Felippo *et al.*, 2021] como conjunto de análise. Tal *corpus* é composto por postagens/*tweets* do domínio do mercado financeiro (mais especificamente sobre ações do índice Ibovespa), retirados da rede social X (antigo *Twitter*). Contabilizam-se 4.517 postagens em português, coletadas em 2014. Segundo os autores, o *corpus* está organizado em função dos seguintes critérios: (i) *Simplificação de código*, como Ausência de hífen; (ii) *Abreviação*, como Contração e Acrônimo; (iii) *Expressão de sentimento*, como Alongamento de pontuação; (iv) *Influência de língua estrangeira*, como Formação verbal; (v) *Marca de oralidade*, como Coloquialismos; (vi) *Elementos metalinguísticos* (do *Twitter*), como Hashtag e Menção; (vii) *Fenômeno do domínio* (Ibovespa), como Ticker e Cashtag.

Para tanto, partiu-se da classificação automática do DANTE-stocks promovida por Ramos e Souza [2024] e Pereira e Souza [2024], que indicaram três classes de textos, a saber postagens bem, mal e medianamente estruturadas, classificadas em função de aspectos semânticos, coesivos e de coerência nas construções linguísticas. A partir dessa classificação, escolheu-se aleatoriamente 180 postagens das três classes identificadas para este estudo preliminar. Em seguida, foi conduzida uma análise semi-automática utilizando a ferramenta rstWeb [Zeldes 2016], em que um anotador sem conhecimentos prévios sobre o mercado financeiro, buscou indicar possíveis relações RST nos *tweets*, bem como seus sinalizadores em potencial, conforme ilustrado na Figura 1.

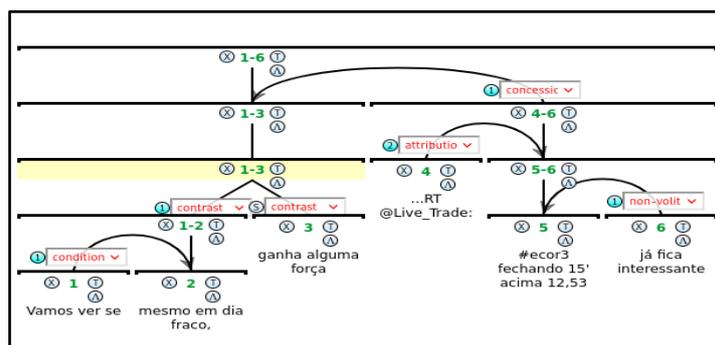


Figura 1. Postagem classificada como “medianamente estruturada”.

Na Figura 1, tem-se um exemplo de postagem anotada com relações RST e seus possíveis sinalizadores. De acordo com a taxonomia proposta por Dantas *et al.* (2024), na relação *Condition*, presente entre as unidades mínimas da construção discursiva, tidas como as *Elementary Discourse Units* (EDUs) 1 e 2, tem-se o “se” como um Marcador Discursivo de subtipo Conjunção. Na relação *Contrast*, entre as EDUs 1-2 e 3, foram anotados “fraco” e “força” como Semântico e de subtipo Antonímia. Na relação *Non-Volitional-Result*, “já fica interessante” foi anotado como Sintático de subtipo oração circunstancial. Por fim, na relação *Concession*, o “#ecor3 fechando 15’ acima 12,53” também foi anotado como sintático de subtipo oração circunstancial. Além disso, foram identificados novos sinalizadores. Na EDU 4, tem-se uma relação de *Attribution* sinalizada por “RT”, que foi anotado com um novo tipo, a saber, Gênero (textual), e “@Live_Trade”, anotado como Sintático de subtipo Citação.

3. Resultados e discussão

No Quadro 1, apresentam-se os sinalizadores em função das relações RST.

Sinalizador		Exemplo	Relação RST
Tipo	Subtipo		
Gênero	Link	(1) @BorisBrianCasoy Petrobras PETR4 Lula capitalizou em 2010 a R\$ 26,30. Últimos 6 meses... Março na bacia das almas! http://t.co/100d2g2AtO	<i>Elaboration</i>
	RT	(2) Vamos ver se mesmo em dia fraco, ganha alguma força ... <i>RT</i> @Live_Trade: #ecor3 fechando 15' acima 12,53 já fica	<i>Attribution</i>
Sintático	Injunção	(3) #ecor3 <i>segue</i> seu caminho rrsr	<i>Elaboration</i>
Semântico	Emoção	(4) #ecor3 segue seu caminho <i>rrsr</i>	Evaluation
	Entidade	(5) @Live_rade MT, tá bem vale5 tb? Grato	
	Citação	(6) @BorisBrianCasoy Petrobras PETR4 Lula capitalizou em 2010 a R\$ 26,30. Últimos 6 meses... Março na bacia das almas! http://t.co/100d2g2AtO	<i>Attribution</i>
Gráfico	Caracter especial	(7) <i>Elétricas lideram altas e baixas com ajuda do governo, Vale renova mínima do ano: Ecorodovia...</i> http://t.co/EkCKzePhxU #infomoney #vale5	<i>Elaboration</i>
	Símbolos	(8) 11/04/14 - 17:18: <i>Maiores Altas:- KROT3 4,17% R\$48,75, CPLE6 3,62% R\$30,94, AEDU3 3,44% R\$13,55, PETR4 2,93% R\$16,14, ALLL3 2,90% R\$8,31.</i>	<i>Elaboration</i>

Quadro 1. Exemplos de novos sinalizadores para textos de CGU.

Do Quadro 1, em (1), o *link* é um sinalizador da relação *Elaboration*, pois indica uma informação extra, que pode ser acessada como complemento da postagem. Em (2), RT

sinaliza a relação *Attribution*, pois indica um “retweet”, ou seja, compartilhamento de uma postagem anterior. Em (3), tem-se *Elaboration* a oração imperativa indica um complemento informacional. Em (4), encontra-se uma *Evaluation*, pois o “rsrsr” indica uma emoção em relação ao que foi dito anteriormente. Em (5), observa-se uma relação de *Attribution*, visto que o autor da postagem menciona “@Live_Trade” e insere na estrutura sintática da postagem, caracterizando uma Entidade. Já em (6), também se tem uma *Attribution*, mas, nesse caso “@BorisBrianCasoy” está sendo citado na postagem, não integrando a estrutura sintática. Em (7), o caracter especial apresenta-se uma *Elaboration*, indicando um complemento à postagem. Por fim, em (8) tem-se relações *Elaboration*, visto que as indicações percentuais e os valores monetários indicam uma informação complementar do que foi citado anteriormente.

Analisando os novos sinalizadores discursivos, nota-se algumas diferenças em relação à tipologia proposta por Dantas *et al.* (2024), como a adição do tipo Gênero e de novos subtipos, como Símbolos, Emoções e Injunção. Com relação aos trabalhos de Taboada e Das (2013) e Liu e Zeldes (2022), há aproximação quanto a proposição do tipo Gênero. Destaca-se ainda que tais trabalhos observaram a combinação entre os sinalizadores para a indicação de relações RST, algo ainda a ser realizado em pesquisas futuras.

4. Considerações finais

Observando as anotações dos textos do *corpus* do CSTNews [Cardoso *et al.*, 2011] e do DANTE-stocks [Di Felippo *et al.* 2021], conclui-se que cada gênero textual pode apresentar características particulares. Nesse sentido, devem ser consideradas tipologias complementares uma à outra.

Os objetivos deste estudo foram alcançados, pois foi investigada a ocorrência de sinalizadores linguísticos-estruturais em *corpora* de textos de CGU com o modelo RST. Para tanto, foram utilizados métodos semiautomáticos para a identificação dos sinalizadores nas relações e, como resultado, apresentou-se uma tipologia complementar com novas categorias de tipo e subtipo de sinalizadores. Em trabalhos futuros, pretende-se ampliar a anotação do *corpus*, verificando possíveis novos sinalizadores, e consequentemente, a atualização da tipologia proposta inicialmente neste trabalho. Além disso, estes dados poderão ser utilizados em outros trabalhos que estejam relacionados ao mercado financeiro, as redes sociais ou qualquer outro ligado a linguística computacional.

Agradecimentos

Este trabalho foi realizado no âmbito do Centro de Inteligência Artificial da Universidade de São Paulo (C4AI -<http://c4ai.inova.usp.br/>), com o apoio da Fundação de Amparo à Pesquisa do Estado de São Paulo (processo FAPESP #2019/07665-4) e da IBM. Este projeto também foi apoiado pelo Ministério da Ciência, Tecnologia e Inovações, com recursos 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. Agradecemos também ao apoio e suporte financeiro do Conselho Nacional de Desenvolvimento Científico e Tecnológico.

5. Referências

- Antonio, J. D. (2017) Mecanismos utilizados pelos destinatários do discurso para identificação de relações de coerência não sinalizadas por conectores. *Delta*, V. 33, pp. 79-108. DOI: <https://doi.org/10.1590/0102-445025798334674077>.
- Cardoso, P. C., Maziero, E. G., Jorge, M. L. C., Seno, E. M., Di Felippo, A., Rino, L. H. M., ... Pardo, T. A. (2011, October). CSTnews-a discourse-annotated corpus for single and multi-document summarization of news texts in Brazilian Portuguese. In *Proceedings of the 3rd RST Brazilian Meeting* (pp. 88-105).
- Dantas, E., Bárbara, L.J.S., Pereira, M.A., Gama, N.S., Almeida, T.J.A., Souza, J.W.C., Cardoso, P.C.F., Rodrigues, R. (2024). *Manual de anotação de sinalizadores discursivos em textos jornalísticos*. São Carlos: Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo. Disponível em <https://repositorio.usp.br/item/003207370>
- Das, D., Taboada, M. (2018). RST Signalling Corpus: A corpus of signals of coherence relations. *Language Resources and Evaluation*, 52, 149-184.
- Di Felippo, A., Postali, C., Ceregatto, G., Gazana, L. S., da Silva, E. H., Roman, N. T., Pardo, T. A. (2021). Descrição preliminar do corpus dantestocks: Diretrizes de segmentação para anotação segundo universal dependencies. In *Anais do XIII Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana* (pp. 335-343).
- Fachada, B. (2019). "Mas" em artigos de opinião: valores e relações retóricas. *elingUP: Revista Eletrônica de Linguística dos Estudantes da Universidade do Porto*, 8(1), p. 108-122.
- Liu, Y., Zeldes, A. (2019). Discourse relations and signaling information: Anchoring discourse signals in RST-DT. *Society for Computation in Linguistics*, 2(1), 314-317.
- Mann, W.C., Thompson, S.A. (1988) Rhetorical Structure Theory: Toward a functional theory of text organization. *Text-interdisciplinary Journal for the Study of Discourse*, v.8, n.3, p. 243-281.
- Pecuch, G. (2021). A relação retórica de elaboração sinalizada pelo marcador discursivo mas em aulas e em entrevistas orais. *Letras Escreve*, 11(1), 43-57.
- Ramos, I.V.M., Souza, J.W.C. (2024). Classificação automática de textos de User-Generated Content utilizando Aprendizagem de Máquina Supervisionado. In *Anais do XV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*. Porto Alegre: SBC.
- Pereira, M.A., Souza, J.W.C. (2024). Subsídios Linguísticos para classificação automática de textos de User-Generated Content. In *Anais do XV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*. Porto Alegre: SBC.
- Rodrigues, R., Souza, J. W., Cardoso, P. C. F. (2023, September). Sinalizadores retórico-discursivos: revisitando a anotação RST no corpus CSTNews. In *Anais do XIV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana* (pp. 249-257).
- Taboada, M., & Das, D. (2013). Annotation upon annotation: Adding signalling information to a corpus of discourse relations. *Dialogue & Discourse*, 4(2), 249-281.
- Zeldes, A. (2016). rstWeb-a browser-based annotation interface for Rhetorical Structure Theory and discourse relations. In *Proceedings of the 2016 Conference of the North*

American Chapter of the Association for Computational Linguistics: Demonstrations (pp. 1-5).

Relações de coerência do português brasileiro: um estudo bibliográfico-documental da RST e seus sinalizadores discursivos

Larissa Jesus Santa Bárbara¹, Roana Rodrigues², Jackson Wilke da Cruz Souza³

¹Universidade Federal de Sergipe (UFS) – São Cristóvão/SE

²Programa de Pós-Graduação em Letras – Universidade Federal de Sergipe (UFS) – São Cristóvão/SE

³Programa de Pós-Graduação em Língua e Cultura - Universidade Federal da Bahia (UFBA) – Salvador/BA

{larissajesus.roana@academico.ufs.br; jackcruzsouza@gmail.com}

Abstract. *This work tends to the development of a bibliographic study (Gil, 2002) in research in RST (Rhetorical Structure Theory), a theory that focuses on the organization of texts, in Brazilian Portuguese (BP) from searches on journals and databases. After the data collection process, the works that are part of this research were organized and described. It is expected to be possible to have a basis and an idea about the works in RST in BP.*

Resumo. *Este trabalho tende para o desenvolvimento de um estudo bibliográfico (Gil, 2002) nas pesquisas em RST (Rhetorical Structure Theory), teoria que tem como foco a organização, entre fragmentos de um texto, no português brasileiro (PB), a partir de buscas em sites de revistas da área e em base de dados. Depois do processo de coleta de dados, foram organizados e classificados os trabalhos que fazem parte desta pesquisa. Espera-se que seja possível ter uma base da condição em que a área se encontra, partindo dos trabalhos acessíveis aqui descritos em RST no PB.*

1. Introdução

A proposta *Rhetorical Structure Theory* – RST [Mann e Thompson 1987], é ser “uma teoria descritiva que tem por objeto o estudo da organização dos textos, caracterizando as relações que se estabelecem entre as partes do texto” [Antonio 2012, p.57]. Trata-se, portanto, de uma proposta de exame da organização de um texto, a partir da análise da coerência textual.

Das e Taboada (2017, p. 1) apontam que “as relações de coerência são frequentemente assinaladas por marcadores discursivos (MDs), que são geralmente considerados como os mais típicos (ou às vezes o único tipo de) sinais no discurso”. Segundo Taboada e Das (2018), os MDs são tidos frequentemente como conectivos, representados majoritariamente por preposições e conjunções. Os autores indicam que utilizar apenas os MDs como identificadores de uma relação RST pode causar duas situações complexas: (i) o número pequeno de relações que podem ser identificadas pelos MDs; e (ii) o uso de um MD pode resultar na confusão entre relações.

Além disso, Antonio (2017) aborda que no modelo RST algumas relações podem ser entendidas com ou sem a presença de um marcador discursivo. Diante disso,

salienta-se que os MDs não são apenas marcadores comuns de um discurso, mas sim um facilitador no momento de identificar e interpretar uma relação.

No contexto do português brasileiro (PB), Rodrigues, Souza e Cardoso (2023) abordaram a necessidade de estudo dos sinalizadores discursivos (SDs). Os SDs podem ser encontrados em diversos tipos de construções linguísticas, facilitando a identificação de outras relações não marcadas por MDs. Cabe pontuar que o que diferencia os MDs dos SDs é o fato de estes poderem características que apenas a de promover conexão, como sinais de natureza fonológica, semântica, sintática e pragmática.

Diante desse cenário, objetivou-se nesta pesquisa realizar um estudo bibliográfico-documental sobre a produção científica centrada na investigação de MDs e SDs em PB. Nosso intuito é contribuir com (i) o fomento de trabalhos centralizados na teoria RST no PB; e (ii) a ampliação da captura de novos sinalizadores que podem ser usados no PB.

2. Pesquisa bibliográfica-documental

Para a realização desta pesquisa bibliográfica-documental [Gil 2002], foram consultadas ferramentas e plataformas de buscas (como Google Acadêmico e Publish or Perish) e, para a realização da descrição aqui proposta, foram utilizados os seguintes critérios de seleção das obras: (i) a maneira em que a teoria RST estava sendo trabalhada; (ii) se a área do texto era focada na linguística/uso na linguagem; e (iii) se os trabalhos focaram no uso da RST no PB.

Nesta seção apresentam-se o trabalhos selecionados nesta pesquisa, bem como alguns dos SDs indicados pelos autores dos trabalhos.

2.1 Proposta de Pardo (2005)

A tese de Pardo (2005) buscou investigar tanto abordagens simbólicas quanto estatísticas para a análise discursiva automática, demonstrando como diferentes tipos de conhecimentos podem contribuir para essa tarefa. Com isso, o autor utilizou o *Corpus TCC*, em que apenas o autor da tese realizou a anotação, sendo essa uma anotação retórica, para evitar qualquer variação nos resultados. Além disso, utilizou a ferramenta *RST Annotation Tool* [Carlson e Marcu 2001], em que é possível realizar a segmentação do texto, escolher as relações retóricas, além de indicar os núcleos e satélites.

Um dos objetivos da tese foi a criação de um analisador discursivo automático chamado DiZer. Além disso, o trabalho centrou nos marcadores discursivos, principais mecanismos linguísticos para a detecção de relações retóricas, bem como em palavras e frases indicativas específicas de gênero e domínio textual.

Ademais, no trabalho do autor ficou marcada a criação de 32 relações da teoria, focados no PB.

2.2 Proposta de Cardoso (2014)

Em sua tese, Cardoso (2014) trabalhou com a sumarização automática de multidocumento, sendo este um de seus objetivos, intensificando, assim, a importância

dos sumários em diversas áreas do conhecimento, principalmente no momento de crescente produção de informações. Além disso, buscava simplificar as informações e ressaltar as que faziam grande impacto no texto.

A partir disso, a autora utilizou duas teorias: CST e RST, sendo esta segunda o foco deste artigo. O uso da teoria RST partiu devido à importância da teoria para os núcleos segmentados, já usada em monodocumentos. Então, Cardoso (2014) propôs realizar isso com uma sumarização de multidocumento, pois facilitaria a sumarização de informações pertinentes no texto. Para pôr em questão a função das teorias com a sumarização, a autora utilizou o *corpus* CSTNews [Cardoso *et al* 2011], que foi anotado em RST, permitindo a análise da estrutura retórica e a relevância da informação.

2.3 Proposta de Antonio (2017)

Essa pesquisa foi realizada para entender o uso e/ou a necessidade das relações RST quando ou não demarcadas por conectivos. Isso fez com que o autor trabalhasse com aspectos relacionados à intencionalidade, semântica, interpretação e entonação. Outrossim, o *corpus* trabalhado foi de língua oral e, por conta dessa condição, os textos passaram por uma retextualização (ou seja, foram transcritos), para que as marcações da conversação pudessem ser preservadas (como, entonação e pausas) na linguagem escrita. Em seguida, esses textos foram entregues para professores do ensino superior, que atuaram como analistas.

Como resultado, as relações foram identificadas por meio de outros sinalizadores, além de MDs. O autor fortaleceu o pressuposto da teoria RST, em que as relações retóricas são de *sentido* e não de *forma*, já que as relações estão presentes no discurso, mesmo na ausência de conectivos. Isso se justifica pelo fato de os informantes pontuarem algumas outras formas de facilitar a identificação das relações, como, por exemplo, tempo verbal, semântica e paralelismo sintático.

2.4 Proposta de Rodrigues, Souza e Cardoso (2023)

O trabalho realizado por Rodrigues, Souza e Cardoso (2023) teve como base as pesquisas de Antonio (2017) e Das e Taboada (2018). Os autores partiram do *corpus* jornalístico CSTNews [Cardoso *et al.* 2011] pré-anotado com o modelo RST e identificaram possíveis sinalizadores das relações retóricas. A escolha de se trabalhar com esse tipo de abordagem é ter uma visão de outras partes do texto que, eventualmente, não foram analisadas.

Os autores fizeram manualmente uma anotação em nove conjuntos de textos, identificando os SDs no texto maior de cada um dos conjuntos, chegando a um rol de sinalizadores para além dos MDs. Ainda, salientam que esses sinalizadores ocorreram de maneira isolada (ou *simples*) ou conjunta (ou *combinada*).

Abaixo, no Quadro 1, apresentam-se exemplos de sinalizadores em função dos trabalhos citados nesta pesquisa. Os trabalhos destacaram ora explícita ora implicitamente os SDs utilizados, com exceção de Cardoso (2014) que não os apresentou.

Quadro 1. Sinalizadores por trabalhos/documentos encontrados.

RELAÇÃO RST	TRABALHOS			
	Pardo (2005)	Cardoso (2014)	Antonio (2017)	Rodrigues, Souza e Cardoso (2023)
<i>Elaboration</i>	Locução verbal	-----	Pontuação (:)	Pontuação (,) + Pronome Relativo (que)
<i>Sequence</i>	MDs (a partir)	----	Tempo Verbal	MD (e) + tempo verbal

A partir do Quadro 1, notou-se que em EDUs anotadas com relações RST *Elaboration* e *Sequence*, por exemplo, ocorreram diferentes SDs e MDs de maneira isolada ou combinada. No trabalho de Antonio (2017), nas unidades anotadas com a relação *Elaboration* ocorreu Pontuação, enquanto, em Rodrigues, Souza e Cardoso (2023) a mesma relação foi identificada pela combinação de dois sinalizadores. Já na relação *Sequence* no trabalho de Pardo (2005) ocorreu uma Locução prepositiva, ao passo que no trabalho de Antonio (2017) o tempo verbal foi apontado como SD dessa relação.

Além do estudo sobre os sinalizadores, Pardo (2005) aplica a RST no desenvolvimento de um dos primeiros analisadores discursivos automáticos para o PB, o DiZer. Já no trabalho de Cardoso (2014), a RST foi aplicada na sumarização automática em PB. Antonio (2017) utiliza a teoria para averiguar, em um *corpus* oral, a possibilidade de existir e, conseqüentemente, identificar uma relação retórica com outros SDs. Por fim, Rodrigues, Souza e Cardoso (2023) buscaram a existência de outros SDs, realizando um estudo a partir de um *corpus* pré-anotado.

3. Conclusões

Nesta pesquisa, buscamos trabalhos científicos que abordassem a RST no PB, sobretudo com relação aos MDs e SDs destacados por eles. Foram encontrados diferentes estudos, entre eles os quatro em destaque. Foi possível perceber a abrangência de perspectivas com relação a esse tema, além da diversidade de áreas e aplicações com que a RST é inserida em português.

Importante destacar que houve limitações sobre o estudo por conta da falta de unificação das bases de dados, o que, eventualmente, sobrepunha os resultados de busca. Assim, cabe, em trabalhos futuros, aplicar outros termos de busca, bem como verificar possíveis sobreposições de resultados entre as bases bibliográficas. Além disso, pretende-se futuramente realizar um aprofundamento nas análises dos sinalizadores discursivos em PB, buscando outras formas de ocorrências e a frequência em diferentes gêneros textuais. Tais resultados poderão subsidiar pesquisas de identificação automática das relações RST, bem como fenômenos discursivos (como desinformação e argumentação) em gêneros textuais e domínios específicos.

Agradecimentos

Este trabalho foi realizado no âmbito do Centro de Inteligência Artificial da Universidade de São Paulo (C4AI -<http://c4ai.inova.usp.br/>), com o apoio da

Fundação de Amparo à Pesquisa do Estado de São Paulo (processo FAPESP #2019/07665-4) e da IBM. Este projeto também foi apoiado pelo Ministério da Ciência, Tecnologia e Inovações, com recursos 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

- Antonio, J. D. (2012) Relações retóricas estabelecidas por orações gerundiais adverbiais. *Alfa: Revista de Linguística*, v. 56, n. 1, p. 55-79.
- Antonio, J. D. (2017) Mecanismos utilizados pelos destinatários do discurso para identificação de relações de coerência não sinalizadas por conectores. *DELTA: Documentação de Estudos em Linguística Teórica e Aplicada*, v. 33, n. 1, p. 79–108.
- Carlson, L.; Marcu, D. (2001) Discourse tagging reference manual. ISI Technical Report ISI-TR-545, v. 54, n. 2001, p. 56.
- Cardoso, P.C.F. *et al.* (2011) CSTNews -A Discourse-Annotated Corpus for Single and Multi-Document Summarization of News Texts in Brazilian Portuguese. [s.l: s.n.]. Disponível em: <<https://sites.icmc.usp.br/taspardo/rst2011-cardosoetal1.pdf> > . Acesso em: 5 out. 2023.
- Cardoso, P.C.F. (2014) Exploração de métodos de sumarização automática multidocumento com base em conhecimento semântico-discursivo. 2014. 182p. Tese (Doutorado). Universidade de São Paulo. Instituto de Ciências Matemáticas e de Computação/Universidade de São Paulo.
- Gil, A. C. (2002) Como elaborar projetos de pesquisa. 4. ed. São Paulo: Editora Atlas.
- Pardo, T. A. S. (2005) Métodos para análise discursiva automática. 2005. Tese (Doutorado em Ciências de Computação e Matemática Computacional) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos.
- Rodrigues, R; Souza, J. W. C; & Cardoso, P. Sinalizadores retórico-discursivos: revisitando a anotação RST no cópús CSTNews. In: Simpósio Brasileiro De Tecnologia da Informação e da Linguagem Humana (STIL), 14., 2023, Belo Horizonte/MG. Anais [...]. Porto Alegre: Sociedade Brasileira de Computação, 2023. p. 249-257.
- Taboada, M.; Das, D. (2013) Annotation upon Annotation: Adding Signalling Information to a Corpus of Discourse Relations. *Dialogue & Discourse*, v. 3, n. 2, p. 249–281.

Subsídios Linguísticos para classificação automática de textos de *User-Generated Content*

Mateus Araújo Pereira¹, Jackson Wilke da Cruz Souza¹

¹ Instituto de Letras (ILUFBA) – Universidade Federal da Bahia (UFBA) – Salvador/BA

² Programa de Pós-Graduação em Língua e Cultura (PPGLinC) Universidade Federal da Bahia (UFBA) – Salvador/BA

pereiramateus@ufba.br, jackcruzsouza@gmail.com

Abstract: *This study aims to classify the structures of User-Generated Content (UGC) texts using the DANTE-stocks corpus, which consists of tweets about the financial market. The texts were manually analyzed and classified based on semantic, cohesive, and coherence criteria according to their linguistic structure, resulting in three classes: (i) well-structured, (ii) moderately structured, and (iii) poorly structured. The integration of these approaches provides a foundation for developing applications in the field of Natural Language Processing related to UGC texts.*

Resumo: *Este estudo visa classificar as estruturas de textos User-Generated Content (UGC), usando o corpus DANTE-stocks, composto por tweets sobre o mercado financeiro. Os textos foram analisados e classificados manualmente com critérios semânticos, coesivos e de coerência em função da estrutura linguística, resultando em três classes: (i) bem, (ii) mediamente e (iii) mal estruturado. A integração dessas abordagens oferece embasamento para o desenvolvimento de aplicações no âmbito do Processamento de Linguagem Natural com relação a textos de UGC.*

1. Introdução

Krumm, Davies e Narayanaswami (2008) definem *User-Generated Content* (UGC), como quaisquer conteúdos criados por usuário em uma plataforma *online*, integrando um conjunto de formatos, como textos, fotos, vídeos, comentários em fóruns ou redes sociais. Essa definição enfatiza o ambiente participativo e colaborativo da UGC, destacando conteúdos, em sua espontaneidade, produzidos por usuários devido à permissividade do ambiente e do gênero textual em que esses textos são produzidos. A UGC é caracterizada por uma variedade de fenômenos linguísticos, como o uso de abreviações, neologismo e estruturas não convencionais, como corrobora Di-Felippo *et al.*, (2021).

A maior parte das pesquisas em Processamento de Linguagem Natural (PLN), até o momento, tem se concentrado na análise de textos formais, em detrimento de textos de UGC. Em virtude da diversidade e informalidade presentes nesses textos, exigem de abordagens direcionadas a compreender suas características únicas. Para que seja possível realizar análises linguísticas considerando particularidades do gênero textual e da linguagem em textos de UGC, objetivamos neste trabalho propor classes que levem em conta aspectos semânticos e estruturais. Nesse sentido, serão

apresentados os resultados do estudo piloto de classificação do *corpus* [Di-Felippo *et al.* 2021]. O estudo teve foco na identificação e categorização dos *tweets* em três classes, a saber: *bem estruturado*, *mediamente estruturado* e *mal estruturado*. Ademais, detalharemos a metodologia aplicada, as abordagens linguísticas adotadas durante o processo, e os resultados preliminares alcançados.

Para tanto, utilizou-se como conjunto de dados o *corpus* DANTE-stocks [DiFelippo *et al.*, 2021], que é composto por aproximadamente 6,700 *posts/tweets* extraídos da plataforma X/Twitter, pertencentes ao domínio do mercado financeiro. Esses textos são caracterizados por uma linguagem curta e direta, contendo jargões e observações específicas ao contexto de uso e suportes midiáticos e **abordam** informações sobre ações, investimentos, notícias econômicas e análises de mercado das bolsas de valores.

Este trabalho está organizado em 3 seções, além desta introdução e da consideração final. Na seção 2, apresentamos uma visão sobre a literatura com estudos que abordam a análise da linguagem em textos de UGC, destacando seus fenômenos linguísticos específicos do gênero textual. Na seção 3, detalhamos a metodologia utilizada na classificação dos *tweets* e quais critérios foram considerados nesse processo. Na seção 4, exploramos as duas abordagens linguísticas aplicadas à classificação dos *tweets* e sua contribuição para a determinação dessas classes. Por fim, na consideração final, discutimos a importância de uma reflexão linguística que vá além das normas cultas da língua e propomos direções para trabalhos futuros.

2. Trabalhos relacionados

Diversos estudos abordam a análise da linguagem em textos de UGC, na perspectiva da caracterização de fenômenos linguísticos específicos, como abreviações, neologismos e estruturas não convencionais. Tagg (2012) investigou a linguagem em *SMS* e redes sociais, observando como a espontaneidade e a informalidade influenciam a estrutura e o conteúdo comunicacional. A autora argumenta que os usuários adaptam os textos às redes sociais, utilizando abreviações e *emojis*, por exemplo, para transmitir mensagens de maneira rápida e demonstrar emoção.

No mesmo contexto, Androutsopoulos (2011) analisou a variedade linguística e a inovação lexical em interações *online*, evidenciando a dinâmica criativa da linguagem digital. O autor pontua que as redes sociais fornecem um espaço para experimentação linguística, permitindo aos usuários criarem vocabulários e formarem novas expressões. Diante disso, é crucial expor a importância de considerar o contexto comunicativo na análise linguística de textos de UGC, que desafiam as normas da GT.

Além desses estudos, Eisenstein (2013) aplicou técnicas de PLN para analisar grande quantidade de textos de UGC, concentrando-se na identificação de padrões linguísticos e fonológicos presentes nesses textos. Eisenstein destaca que a escrita em redes sociais pode refletir o modo de pronúncia das palavras, com a substituição de letras para representar sons específicos. O autor argumenta que a escrita, nesse contexto, tem aspectos fonológicos, ao demonstrar uma conexão entre a fala e a escrita.

É importante salientar que essas pesquisas demonstram como abordagens computacionais podem ser usadas para perceber a diversidade e a singularidade da linguagem em ambientes digitais. Dessa forma, os resultados dos trabalhos contribuem para o desenvolvimento de modelos de PLN para lidar com a variedade linguística presente em textos de UGC.

Somado a isso, estudos anteriores [Tagg 2012; Eisenstein 2013], focam nos fenômenos linguísticos como abreviações e a inovação lexical nas redes sociais, mas poucos se dedicam a classificar esses textos quanto a sua estrutura. Assim, esta pesquisa é um dos poucos estudos a propor uma categorização de *tweets* baseando-se em critérios de coesão, coerência e semântica. Trabalhos mais recentes [Longaretti 2021; Steinhauser e Botassini 2023] indicam a ocorrência de fenômenos linguísticos voltados aos níveis semântico, gramatical, lexical, discursivo e prosódico em redes sociais.

3. Metodologia

Este estudo passou por duas etapas metodológicas. Na primeira, foi conduzido um estudo exploratório nos *tweets*, com ênfase nos aspectos morfossintáticos - a análise da estrutura e organização das palavras dentro das frases, como combinações de artigo + substantivo + verbo. Para tal, foram analisadas 70 postagens selecionadas aleatoriamente do *corpus* DANTE-stocks [Di Felippo *et al.* 2021], com o objetivo de identificar padrões estruturais característicos. Apesar disso, os resultados não evidenciaram regularidade, chegando-se a 70 possibilidades de combinações. Nesse contexto, na segunda etapa definiram-se três classes para os *tweets*, baseadas em aspectos semânticos (compreensão da mensagem), coesão (conexão entre elementos linguísticos do texto) e coerência (organização lógica dos itens linguísticos).

Dada a inviabilidade de propor um padrão morfossintático para cada *tweet* do *corpus*, foram propostas três classes que consideram a estrutura linguística de cada um deles. Para tanto, baseamo-nos em critérios de naturezas semântica, coesiva e coerente para cada uma das classes propostas, como demonstrado no Quadro 1.

CRITÉRIOS	CLASSES		
	Bem estruturado	Mediamente estruturado	Mal estruturado
Semântica Compreensão textual	Postagem totalmente compreensível	Postagem parcialmente compreensível, mas requer certo conhecimento do domínio	Postagem totalmente incompreensível.
Coesão Conexão entre proposições	Boa conexão.	Conexão imprecisa, apesar de presente	Pouquíssima ou nenhuma conexão
Coerência Organização lógica	Boa organização lógica	Organização lógica imprecisa	Carência e/ou ausência de organização lógica

Quadro 1. Descrição dos critérios entre as classes.

No Quadro 1 descrevem-se os critérios usados para categorizar os *tweets* nas três classes propostas. *Tweets bem estruturados* têm alta compreensão semântica e coesiva, mesmo sem conhecimento prévio do contexto ou domínio. *Tweets mediamente estruturados* têm semântica parcialmente compreensível e coesão limitada, necessitando de algum conhecimento prévio. Por fim, *tweets mal estruturados* têm baixa compreensão semântica, conexões fracas e falta de lógica, dificultando a interpretação do conteúdo.

A classificação dos *tweets* foi realizada manualmente por três anotadores, sem que tivessem acesso à anotação individual do grupo. Ao fim desse processo os resultados foram comparados automaticamente. Para tanto, quando havia concordância total sobre a classificação entre os três anotadores, o *tweet* era rotulado diretamente com a classe indicada. Porém, quando houve casos de discordância, utilizou-se da regra da

maioria: a classe com maior voto entre os anotadores predominava como a escolha final para a rotulação do *tweet*.

4. Resultados e Discussões

Como resultado, foram analisados 180 *tweets*, distribuídos entre as classes *bem estruturada* (82 *tweets*), *mediamente estruturada* (59) e *mal estruturada* (39). Essa etapa envolveu a leitura cuidadosa de cada *tweet*, considerando os aspectos semânticos e discursivos para essa classificação. Após a determinação dos critérios de cada uma das classes, os *tweets* foram classificados manualmente, como demonstrado no Quadro 2.

<i>TWEETS</i>	CLASSIFICAÇÃO
1) PETR4 subiu na bolsa 13,50. Muito bem, estou surpreso com o resultado.	Bem estruturado
2) Ano passado eu falei que até o final de 2104 #PETR4 estaria abaixo de R\$ 10,00 mas acho que errei, não vai demorar tanto.	Bem estruturado
3) vai, oi4. um trouxe... ops... investidor precisa pagar as minhas férias 4) que linda era esa mina chabonnn	Mediamente estruturado
5) <,Alexander Cruz3 *-* 6) @victoriabril_ forra contesta	Mal estruturado

Quadro 2. Exemplos de *tweets* e sua classificação.

No Quadro 2, o *tweet* em (5), classificado como *Mal estruturado*, não permite compreensão sobre o conteúdo, pois há apenas a indicação de um nome e a inclusão de um *emoji* (“ *-* ”), tido como um recurso paralinguístico, desempenhando uma função discursiva. Já o conteúdo de (1), classificado como *Bem estruturado*, é compreensível por si só, se enquadrando nos três critérios: a semântica é direta e acessível, a coesão é mantida entre os elementos linguísticos, e a coerência permite uma organização lógica que contribui o entendimento do texto como um todo. Por fim, (3), classificado como *mediamente estruturado*, apresenta uma sequência lógica do conteúdo, mas que apresenta elementos que podem prejudicar a estruturação do conteúdo, como a interrupção da fala (em “um trouxe...”) e a correção do que foi dito (indicado por “ops...”).

5. Considerações finais

Neste estudo procuramos delimitar critérios para classificar *tweets* em função da estrutura de composição e compreensão do conteúdo. Esses métodos poderão ser utilizados para o treinamento e teste de algoritmos de Aprendizado de Máquina para gerar classificadores automáticos, permitindo escalabilizar o processo. Proporcionando uma compreensão sobre os aspectos comunicativos em plataformas digitais. A integração de abordagens linguísticas e computacionais na análise de *tweets*, representa um avanço na capacidade de compreender e modelar a comunicação digital, oferecendo novas ferramentas para a análise de grandes volumes de dados textuais e melhorando a interação entre humanos e máquinas em ambientes de redes sociais.

É importante destacar a limitação de perspectivas prescritivistas e normativas, alinhadas à Gramática Tradicional nesse tipo de pesquisa. Eventualmente, *tweets* que foram previamente classificados como *mediamente estruturados* poderiam ser classificados como *bem estruturados*, como a postagem (2) do Quadro 1, por exemplo. Assim, em análises futuras será imprescindível considerar outras perspectivas

linguísticas que levem em conta o uso da língua, como a Gramática Funcional, compreendendo limites e contribuições metodológicas e teóricas mútuas.

Por fim, destaca-se a importância deste trabalho no escopo de anotação de *corpus*. O resultado deste trabalho permitirá identificar quais *tweets* possuem diferentes características estruturais e linguísticas que poderão ser úteis para a identificação de fenômenos discursivos, por exemplo. Isso permitirá, portanto, a aplicação de modelos teóricos nesse sentido, avançando o estado da arte em PLN para o PB.

Agradecimentos

Este trabalho foi realizado no âmbito do Centro de Inteligência Artificial da Universidade de São Paulo (C4AI -<http://c4ai.inova.usp.br/>), com o apoio da Fundação de Amparo à Pesquisa do Estado de São Paulo (processo FAPESP #2019/07665-4) e da IBM. Este projeto também foi apoiado pelo Ministério da Ciência, Tecnologia e Inovações, com recursos 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. Além disso agradecemos ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), a Fundação de Amparo à Pesquisa do Estado da Bahia (FAPESB) pelo financiamento e suporte.

Referências

- Androutsopoulos, J. (2011). Language Change and Digital Media: A Review of Conceptions and Evidence. *Standard languages and language standards in a changing Europe*, 1, pp.145-159.
- Di-Felippo *et al.* (2021). Descrição Preliminar do Corpus DANTEStocks: Diretrizes de Segmentação para Anotação segundo Universal Dependencies. In *Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana (STIL)*, (pp. 335-343). Porto Alegre: SBC. DOI: <https://doi.org/10.5753/stil.2021.17813>
- Eisenstein, J. (2013). Phonological Factors in Social Media Writing. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, (pp.11-19). Atlanta: Association for Computational Linguistics.
- Krumm, J.; Davies, N. e Narayanaswami, C. (2018) User-generated content. *IEEE Pervasive Computing*, v. 7, n. 4, pp.10-11.
- Longaretti, R. B. (2021). *O difícil de dizer em texto de instrução ao sócio sobre o trabalho docente: uma análise de fenômenos linguísticos prosódicos*. 163f. Dissertação (Mestrado em Letras) – Programa de Pós-Graduação em Letras – Universidade Tecnológica Federal do Paraná, Pato Branco, Paraná, 2021.
- Steinhauser, V. L. F., e Botassini, J. O. M. (2023). Vem sabonetar aqui fora! Um estudo multissistêmico do verbo Sabonetar sob influência do twitter e dos reality shows. *Papéis: Revista do Programa de Pós-Graduação em Estudos de Linguagens - UFMS*, 27(53), pp.114-138.
- Tagg, C. (2012). *Discourse of Text Messaging: Analysis of SMS Communication*. London: Continuum.

Relações de coerência do espanhol peninsular: Um estudo bibliográfico-documental da *Rhetorical Structure Theory*

Ewerson Dantas¹, Roana Rodrigues², Jackson Wilke da Cruz Souza³

¹Universidade Federal de Sergipe (UFS) – São Cristóvão/SE

²Programa de Pós-Graduação em Letras – Universidade Federal de Sergipe (UFS) – São Cristóvão/SE

³Programa de Pós-Graduação em Língua e Cultura - Universidade Federal da Bahia (UFBA) – Salvador/BA

roana@academico.ufs.br; {ewersonndantad, jackcruzsouza}@gmail.com

Abstract. *This work aims to develop a bibliographical study [Gil,2002] of academic works on RST (Rhetorical Sistematic Theory), an analytical discourse theory focused on the organization and analysis of texts in peninsular Spanish (ESP). After collecting texts from international journals and databases, we proposed an organization and classification of the works, based on how RST is used.*

Resumo. *Esse trabalho visa o desenvolvimento de um estudo bibliográfico [Gil 2002] de trabalhos acadêmicos em RST (Rhetorical Sistematic Theory), teoria discursivo-analítica focada na organização e análise de textos, no espanhol peninsular (ESP). Após a coleta de textos, a partir de revistas internacionais e bases de dados, foi proposta uma organização e classificação dos trabalhos, tendo como base a forma como a RST é utilizada.*

1. Introdução

A RST (*Rhetorical Sistematic Theory*) é uma teoria discursivo-analítica desenvolvida, a princípio, para análise de textos por Mann e Thompson (1987), sendo muito utilizada em pesquisas de Processamento de Língua Natural (PLN), no campo da Linguística Computacional. O principal ponto dessa teoria é o estudo da organização do texto, tendo em vista a coerência estabelecida nele, utilizando dos marcadores discursivos para esse estudo. Nesse sentido a RST é vista em diversos tipos de pesquisas e trabalhos dentro da PLN, para exemplificação, temos a criação do DiSeg [Da Cunha *et al.* 2010], um segmentador discursivo para língua espanhola, como também dos diversos trabalhos comparatistas entre espanhol e basco feitos por Da Cunha e Iruskieta.

Neste trabalho, o RST é o principal tema para o levantamento bibliográfico [Gil 2002] realizado com pesquisas dentro do espanhol peninsular (ESP), ou seja, aquele falado na Espanha. Após a coleta dos textos, a partir de revistas internacionais e ferramentas de bases de dados, foi feita uma organização e classificação destes a partir da observação da forma como a teoria é trabalhada.

A organização do artigo se dará da seguinte forma, além desta Introdução: na Seção 2 será abordada a metodologia utilizada para a pesquisa dos trabalhos em RST no espanhol peninsular; na Seção 3 os resultados obtidos serão relatados com foco na

exposição da classificação feita a partir do levantamento bibliográfico; na Seção 4 constam as considerações finais do trabalho.

2. Metodologia

O primeiro passo para o levantamento bibliográfico foi a procura por trabalhos de pesquisa que incluíssem a RST dentro do espanhol peninsular, para isso foram utilizadas duas plataformas de pesquisa: o Google Acadêmico e o *ACL Anthology*. Os termos de buscas utilizados foram “RST”, “spanish” e “español”. Como resultado foram encontrados 2330 resultados no Google acadêmico, e 601 resultados no *ACL Anthology*, totalizando 2931 trabalhos. A partir dos resultados encontrados houve uma separação entre aqueles que realmente se enquadram como objeto de estudo deste trabalho, seguindo os seguintes critérios; (i) a forma como a RST era citada no resumo e nas palavras-base do trabalho, (ii) se os criadores da teoria, Mann e Thompson, estavam citados na referência, (iii) se o trabalho realmente focava na língua espanhola, utilizando dos resumos e introduções dos textos como base. Ao final desta separação, 30 trabalhos se enquadraram nos requisitos citados.

Após essa separação, uma segunda fase foi feita focando em coletar somente textos que trabalhassem com o espanhol peninsular, para isso duas coisas foram observadas: (i) a instituição a qual o trabalho está ligado, e (ii) os pesquisadores responsáveis pela sua execução. Assim, ao final, somente textos que trabalham com a RST dentro da área da PLN, e tem como objeto de estudo o espanhol peninsular, permaneceram para análise.

3. Resultados e discussão

Partindo da pesquisa realizada, chegou-se a sete trabalhos que cumpriam todos os requisitos apresentados. Em seguida, foi proposta uma classificação em função da maneira como a RST é abordada nos trabalhos. Chegou-se, então, em duas categorias, a saber: “Estudos comparativos”, que são pesquisas que compararam duas ou mais línguas, ou ainda gêneros textuais, observando estruturas e relações de coerência mais recorrentes; e “Estudos em interface com PLN”, que são trabalhos mono e multilíngue centrados na criação de ferramentas e recursos em Linguística computacional.

3.1. Estudos comparativos

a) Trabalho de Da Cunha e Iruskieta (2010)

Neste estudo os autores utilizaram a RST para uma análise comparativa entre o Espanhol e o Basco, tendo como foco as ocorrências de relações retóricas, como também seus respectivos marcadores. Para tanto, realizaram a análise das relações RST e seus marcadores a partir de 20 resumos retirados do periódico científico *Gaceta Médica de Bilbao*, que tem como domínio a medicina. Após a coleta dos textos, houve mais duas fases metodológicas: a fase *quantitativa*, feita a partir da contabilidade dos aspectos discursivos do *corpus*; e a fase *qualitativa*, focada na análise da ambiguidade dos marcadores e na forma como eles refletem cada relação. Os autores observaram um número maior de marcadores discursivos no Basco, além de notarem similaridade na quantidade de relações retóricas entre as duas línguas analisadas.

b) Trabalho de Iruskietea e Da Cunha (2010)

Neste trabalho os autores usam a RST em um estudo comparatista entre domínios distintos (Medicina e a Terminologia) de textos produzidos em Espanhol e Basco, e publicados na *Gaceta Médica de Bilbao* e no Congresso Internacional de Terminologia de 1997. O objetivo dos autores foi utilizar as relações retóricas como um meio para caracterizar os domínios. A análise dos textos foi feita a partir do processo de anotação do *corpus*, seguido pela análise discursiva, realizada para delimitar as diferenças entre as anotações entre os dois idiomas. Por fim, foi feita a análise quantitativa, focada na quantidade de relações ocorridas em cada área. Os autores notaram pontos interessantes tanto na comparação *entre áreas* (como o maior aparecimento da relação *Result* no *corpus* de Medicina, e de *Interpretation*, no *corpus* de Terminologia) e *entre as línguas* (diferenças na pontuação e sintáticas, principalmente pelo Basco ser uma língua aglutinante, onde a maioria dos morfemas está junto a palavra).

c) Trabalho de Iruskietea, Da Cunha e Taboada (2014)

Neste trabalho os autores fizeram uso de três diferentes idiomas (Inglês, Espanhol e Basco), buscando comparar a estrutura retórica das línguas a partir da anotação e análise de *corpus* dos três autores/anotadores. O *corpus* trabalhado continha textos da Conferência Internacional de Terminologia, realizada no ano de 1997. Os autores analisaram 15 resumos que continham as três línguas. A anotação semi-automática foi feita a partir da ferramenta RSTTool [O'Donnell 2000]. Ao final, os autores chegaram à conclusão de que o par Inglês-Espanhol continha o maior grau de concordância entre os anotadores, seguido pelo Espanhol-Basco, e por último Inglês-Basco, sendo estas últimas as línguas com menor concordância entre si porque estão tipologicamente mais distantes, além de não ter contato próximo como o par Espanhol-Basco.

3.2. Estudos em interface com PLN

a) Trabalho Da Cunha et al. (2010)

Os autores propuseram um segmentador automático de unidades mínimas de análise em RST, tidas como *Elementary Discourse Units* (EDU). Tais unidades podem ser frases ou orações, a partir das quais se constroem as árvores discursivas [Tofiloski, Brook e Taboada 2009]. Como resultado, desenvolveram o DiSeg, testando-o em um *corpus gold standart*, composto por artigos retirados da *Gaceta Médica de Bilbao*. Após testes e análises realizados, os autores destacam que a ferramenta DiSeg teve um bom desempenho ao segmentar EDU, especialmente quando comparado a outros segmentadores encontrados fora da língua espanhola. Por fim, os autores apontaram que os erros de segmentação se concentraram diante da partícula “y” quando aparecia antes da partícula “que”.

b) Trabalho de Da Cunha, Torres-Moreno e Sierra (2011)

Nesta pesquisa o objetivo é desenvolver um *corpus* anotado em RST para o Espanhol (*RST Spanish Treebank*) visando ao desenvolvimento de um analisador discursivo. Para a formação do *corpus*, os autores objetivaram a diversidade de gêneros textuais (como artigos científicos, teses de doutorado e livros didáticos, por exemplo) e domínios (como Astrofísica, Engenharia, Economia e Direito, por exemplo). A anotação RST foi feita com o RSTTool, por uma equipe de 10 anotadores com diferentes graus de formação

acadêmica. Os autores evidenciaram que o *RST Spanish Treebank* é um recurso gratuito, desenvolvido para melhoria dos estudos da teoria, contando inclusive com uma interface *online* [Da Cunha *et al.* 2011]. Ao final, os autores também ressaltaram a necessidade de melhorias no processo de anotação e no *corpus*, como a necessidade de se incluir mais anotadores, a utilização de diferentes medidas de concordância, além do aumento do número de texto.

c) Trabalho de Cao, Da Cunha e Iruskieta (2018)

Os autores objetivaram a criação do primeiro *corpus* paralelo entre o Espanhol e o Chinês anotado em RST, o *Spanish-Chinese Treebank*. Os textos que compunham o *corpus* são de diferentes gêneros textuais (como resumos de trabalhos acadêmicos, notícias e anúncios), com diferentes quantidades de palavras. A ferramenta utilizada para anotação dos textos foi a RSTTool, e o grupo de anotadores era bilíngue ou falantes nativos de somente uma das línguas. Ao fim, o *Spanish-Chinese Treebank* obteve uma anotação com alta concordância, podendo ser um recurso a ser utilizado no desenvolvimento de ferramentas (como tradutores automáticos) ou abordagens (como Aprendizado de Máquina) em PLN. No âmbito do Espanhol e Chinês.

4. Considerações finais

Neste trabalho foi desenvolvido um levantamento bibliográfico e documental acerca de pesquisas acadêmicas que abordassem a teoria RST tendo como escopo o Espanhol peninsular. A partir do levantamento realizado foram encontrados seis trabalhos que se enquadraram em todos os requisitos impostos nesta pesquisa, sendo a pesquisadora Iria Da Cunha colaboradora em todos eles.

As análises feitas a partir destes textos mostraram um interesse na evolução da teoria RST no Espanhol, destacado pelos trabalhos com foco no desenvolvimento de ferramentas e recursos em PLN para o idioma. Demais estudos demonstram uma tendência aos estudos comparatistas em conjunto com o ESP, tanto com línguas também faladas na Espanha (Basco), como com línguas mais distantes geograficamente (Inglês e Chinês).

Tais observações corroboram os apontamentos feitos por Rodrigues, Souza e Cardoso (2023). Os autores destacaram a importância dos estudos centrados em segmentação, anotação e análise comparativa entre línguas, reforçando inclusive a ideia entre estudos descritivos e comparativos utilizando o Português e o Espanhol, em suas mais diversas variações, aspecto que é fortalecido pela dinâmica político-linguística na América do Sul.

Importante ressaltar que é possível encontrar outros resultados a partir de diferentes palavras chaves utilizadas no momento das pesquisas e buscas por textos. Sendo assim, trabalhos futuros poderão utilizar dessa lacuna para novas revisitações à bibliografia da RST no Espanhol peninsular.

Agradecimentos

Este trabalho foi realizado no âmbito do Centro de Inteligência Artificial da Universidade de São Paulo (C4AI -<http://c4ai.inova.usp.br/>), com o apoio da Fundação de Amparo à Pesquisa do Estado de São Paulo (processo FAPESP #2019/07665-4) e da IBM. Este

projeto também foi apoiado pelo Ministério da Ciência, Tecnologia e Inovações, com recursos 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. Além disso agradecemos ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) pelo financiamento e suporte.

5. Referências

- Cao, Shuyuan, Da Cunha, Iria, e Iruskieta, Mikel. (2018) The rst spanish-chinese treebank. In: *Proceedings of the joint workshop on linguistic annotation, multiword expressions and constructions*. New Mexico/USA: Association for Computational Linguistics. p. 156-166. Disponível em: <https://aclanthology.org/W18-49>
- Da Cunha, Iria *et al.* (2011) The RST Spanish treebank on-line interface. In: *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*. Hissar/Bulgaria: Association for Computational Linguistics. p. 698-703. Disponível em: <https://aclanthology.org/R11-1000>
- Da Cunha, Iria *et al.* (2012) DiSeg 1.0: The first system for Spanish discourse segmentation. *Expert Systems with Applications*, v. 39, n. 2, p. 1671-1678. Disponível em: <https://hal.science/hal-01314824>
- Da Cunha, Iria, e Torres-Moreno, Juan-Manuel; e Sierra, Gerardo. (2011) On the development of the RST Spanish Treebank. In: *Proceedings of the 5th Linguistic Annotation Workshop*. Oregon/USA: Association for Computational Linguistics. p. 1-10. Disponível em: <https://aclanthology.org/W11-0400>
- Gil, Antônio Carlos. (2002) *Como elaborar projetos de pesquisa*. Editora Atlas SA.
- Iruskieta, Mikel; e Da Cunha, Iria (2010) *Marcadores y relaciones discursivas en el ámbito médico: un estudio en español y euskera*. Vigo: Universidade de Vigo. p. 146-159. Disponível em: http://ixa.si.ehu.es/sites/default/files/dokumentuak/3965/AESLA_marcadores.pdf
- Iruskieta, Mikel; e Da Cunha, Iria. (2010) El potencial de las relaciones retóricas para la discriminación de textos especializados de diferentes dominios en euskera y español. *Calidoscópico*, v. 8, n. 3, p. 181-202. Disponível em: <https://www.redalyc.org/articulo.oa?id=571561875003>
- Iruskieta, Mikel; e Da Cunha, Iria; e Taboada, Maite. (2015) A qualitative comparison method for rhetorical structures: identifying different discourse structures in multilingual corpora. *Language resources and evaluation*, v. 49, p. 263-309. DOI: <https://doi.org/10.1007/s10579-014-9271-6>
- Mann, William C., e Thompson, Sandra (1987) *A. Rhetorical structure theory: Description and construction of text structures*. In: *Natural language generation: New results in artificial intelligence, psychology and linguistics*. Dordrecht: Springer Netherlands. p. 85-95.
- O'Donnell, Michael. (2000) RSTTOOL 2.4-A markup tool for rhetorical structure theory. In: *INLG'2000 Proceedings of the First International Conference on Natural Language Generation*. Mitzpe Ramon/Israel: Association for Computational Linguistics. p. 253-256. Disponível em: <https://aclanthology.org/W00-14>

Rodrigues, R., Souza, J., e Cardoso, P. (2023). Sinalizadores retórico-discursivos: revisitando a anotação RST no corpus CSTNews. In *Anais do XIV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*. Porto Alegre: SBC. p. 249-257. DOI: <https://doi.org/doi:10.5753/stil.2023.234120>

Tofiloski, Milan; e Brooke, Julian; e Taboada, Maite. (2009) A syntactic and lexical-based discourse segmenter. In: *Proceedings of the ACL-IJCNLP 2009 conference short papers*. Suntec/Singapore: Association for Computational Linguistics. p. 77-80. Disponível em: <https://aclanthology.org/P09-2020>

Estratégias automáticas para análise da concordância da anotação de Sinalizadores Discursivos

Gabriel Sizinio Bomfim Cruz¹, Jackson W. C. Souza², Paula C. F. Cardoso³

¹Instituto de Computação – Universidade Federal da Bahia, Salvador/BA

²Programa de Pós-Graduação em Língua e Cultura (PPGLinC) – Universidade Federal da Bahia (UFBA) – Salvador/BA

³Faculdade de Computação – Universidade Federal do Pará, Belém/PA

{gsizinio, jackcruzsouza}@gmail.com, pcardoso@ufpa.br

Abstract. *The process of annotating a corpus using Rhetorical Structure Theory (RST) has very clear and defined steps, among which the analysis of agreement between annotators stands out. In this work we present two agreement analysis strategies (gold and silver) based on the Krippendorff Alpha measure. The results point to advanced advances in this type of analysis and the possibility of replication by other works in this segment.*

Resumo. *O processo de anotação de um corpus utilizando a Teoria da estrutura retórica (RST) possui etapas bem claras e definidas, dentre as quais destaca-se a análise da concordância entre os anotadores. Neste trabalho apresentamos duas estratégias de análise da concordância (gold e silver) com base na medida de Krippendorff Alpha. Os resultados apontam significativos avanços para esse tipo de análise e a possibilidade de replicação por outros trabalhos nesse segmento.*

1. Introdução

A *Rhetorical Structure Theory* (RST) [Mann e Thompson, 1987] é uma teoria linguística, amplamente utilizada no Processamento de Linguagem Natural (PLN) para análise de discursos. Tal modelo teórico se propõe a descrever a organização estrutural dos textos com base nas relações retóricas que ocorrem entre suas partes. Essas relações, como *Justify*, *Condition* e *Elaboration*, por exemplo, são capazes de evidenciar interpretações e intenções por meio de como diferentes fragmentos do texto se relacionam entre si. Majoritariamente, essas relações têm sido identificadas por meio de marcas explícitas na superfície textual, como Marcadores discursivos (preposições e conjunções) e outros Sinalizadores discursivos (como pontuação e sentido do verbo).

A identificação dessas marcas é feita predominantemente por meio de anotação de *corpus*. Hovy e Lavid (2010) destacam que esse processo possui etapas bem definidas e, dentre elas, são fundamentais as etapas de análise da anotação e avaliação do nível de concordância entre os anotadores, já que, a depender da concordância obtida, o processo de anotação do *corpus* pode seguir para outras etapas. Ainda, segundo os autores, a baixa concordância entre os anotadores indica que não há consistência no trabalho para permitir que os algoritmos de aprendizado de máquina (AM) sejam treinados a partir desse material. Por outro lado, uma concordância alta indica que o processo pode prosseguir para a anotação de uma maior parte do *corpus* e, conseqüentemente, o material obtido ser usado para treinamento de sistemas baseados em AM. Nesse cenário, dado que análise da concordância é uma questão central para garantir a qualidade dos dados anotados e o prosseguimento do processo de anotação, é necessário garantir que, em sua automação, os resultados possam ser confiáveis.

Diante disso, nosso objetivo neste trabalho é apresentar estratégias automáticas para a medição e análise da concordância entre humanos na tarefa de anotação de Sinalizadores Discursivos (SDs) para as relações retóricas da RST. Mais especificamente, propomos e comparamos duas abordagens distintas para análise da concordância, uma mais restrita (*Gold*) e outra mais flexível (*Silver*), avaliando o alcance de ambas.

Para tanto, este artigo está organizado em quatro seções, além desta Introdução. Na Seção 2 é apresentada a metodologia aplicada para a medição e análise da concordância entre os anotadores e na Seção 3, os resultados obtidos a partir dela. Por fim, na Seção 4, tecemos considerações finais acerca deste trabalho.

2. Metodologia

Este trabalho se concentra na medição e avaliação da concordância entre anotadores na tarefa de marcar SDs em textos jornalísticos. A anotação foi feita semiautomaticamente por Cardoso *et. al* (2024) em uma amostra do *corpus* CSTNews [Cardoso *et. al* 2011], em que cada texto, já com as relações RST, teve a indicação de SDs por três anotadores diferentes. A anotação foi realizada a partir da ferramenta rstWeb [Zeldes 2016], seguindo as diretrizes propostas por Dantas *et. al* (2024).

A partir da anotação de SDs realizada, propôs-se um *pipeline* de pré-processamento e construção de um algoritmo para analisar a concordância entre os anotadores. Este *pipeline* envolveu a eliminação de inconsistências nas anotações e a ordenação dos *tokens* indicados pelos anotadores.

A análise da concordância da anotação pode ser subdividida em (i) escolha da medida de análise, (ii) identificação dos *tokens* anotados e (iii) comparação entre as anotações, sobretudo a comparação entre as escolhas dos *tokens* feitas pelos anotadores. Uma medida comumente utilizada é a *Cohen Kappa* [McHugh 2012]; porém, no escopo deste trabalho, utilizá-la criaria a limitação de comparar as anotações somente entre dois anotadores. Como a anotação de cada texto foi realizada por três anotadores diferentes, optou-se pela medida *Krippendorff Alpha* [Krippendorff 2011]. Tal medida estatística foi utilizada para avaliar a confiabilidade entre dois ou mais anotadores, apresentando mais robustez ao trabalhar com dados categóricos e intervalares, como é o caso das anotações de SDs. Seu valor varia de -1 a 1, em que valores próximos a 1 indicam alta concordância entre os anotadores, valores próximos a 0 indicam baixa concordância, e valores próximos a -1 indicam discordância sistemática.

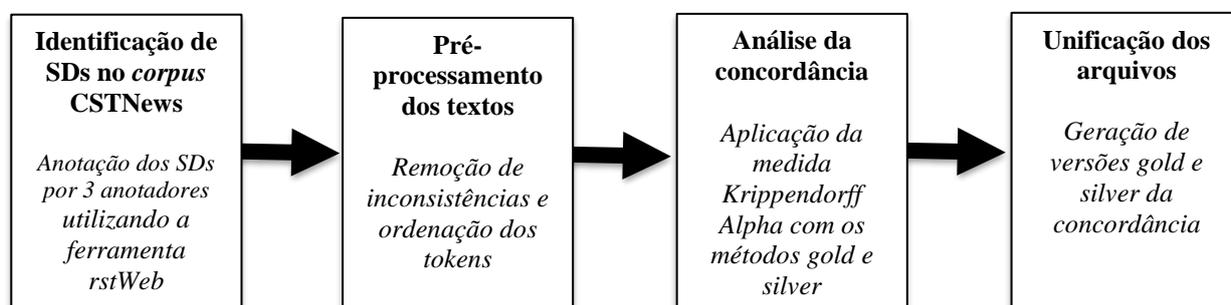


Figura 1. Síntese do processo de anotação e análise da concordância.

Para tanto, implementamos esse cálculo a partir da biblioteca do *Natural Language Toolkit* (NLTK) [Bird, Klein e Loper 2008]. O cálculo foi feito a partir de arquivos .rs3, gerados pela ferramenta rstWeb. Esse arquivo, que é baseado em .xml, está organizado a partir de *tags*, em que cada uma delas se refere a algum aspecto da anotação ou do texto. Nesta pesquisa, as

tags analisadas foram do tipo *signal*, referentes aos SDs indicados por cada um dos sinalizadores em relação ao *Elementary Discourse Unit* (EDU)¹.

3. Resultados

Para análise dos resultados, foram propostas duas análises sobre a concordância. A primeira análise, tida como *gold*, é mais restrita e impõe que os anotadores concordem estritamente na escolha de um *token*, indicando que eles anotaram os mesmos sinalizadores discursivos. Já na segunda, tida como *silver*, é mais flexível, e propõe arbitrariamente a elaboração de uma janela com escopo entre -5 e +5, em relação ao *token*-alvo. Nesta última proposta, os anotadores concordam se um deles anotou um *token* e o outro anotou outro que esteja no espaço da janela de cinco *tokens*, tanto para direita quanto para a esquerda. Na Tabela 1, são apresentados exemplos dessas duas propostas.

Texto	Tokens anotados		Token-alvo	Concordância	
	Anotador A	Anotador B		<i>Gold</i>	<i>Silver</i>
Eu adoro bolo de chocolate, torta de morango e mousse de limão.	“Eu”, “adoro”, “chocolate”	“Eu”, “adoro”, “bolo”, “torta”, “morango”, “mousse”	“chocolate”	“Eu”, “adoro”	“Eu”, “bolo”, “adoro”, “torta”, “morango”

Tabela 1. Exemplo de comparação entre análises *gold* e *silver*.

A partir da Tabela 1 tem-se que a análise *gold* é mais restrita, uma vez que os anotadores A e B, por exemplo, concordam apenas nos *tokens* “Eu” e “adoro”. Já a análise *silver*, por conta da janela de concordância ser mais flexível, em relação ao *token*-alvo “chocolate” (posição 0) foi possível considerar os *tokens* “bolo” (posição -2), “torta” (+2) e “morango” (+4) na análise, pois foram indicados pelo anotador B, e se encontram dentro da janela de concordância em relação à “chocolate”. Embora os anotadores não tenham indicado exatamente os mesmos *tokens*, por conta da proximidade, é possível considerar os tokens de maneira mais flexível.

Para operacionalizar a análise, os dados anotados foram organizados de forma a permitir a comparação entre diferentes anotadores. Cada anotação foi processada e armazenada para que a realização da análise da concordância e obtenção do *Krippendorff Alpha* fosse feita.

Na Tabela 2, tem-se a comparação entre os dois métodos, considerando quatro experimentos. Cada texto foi anotado por três anotadores: um mais experiente e outros dois menos experientes. Nos experimentos I e II foram realizados os cálculos do *Krippendorff Alpha* entre anotadores mais e menos experientes separadamente em relação à tarefa de anotação de *corpus*. No Experimento III foi realizado o cálculo entre os anotadores menos experientes. Por fim, no Experimento IV, calculou-se a concordância entre os três anotadores.

Métodos	Experimentos			
	I	II	III	IV
<i>Gold</i>	0.477	0.433	0.455	0.455
<i>Silver</i>	0.628	0.680	0.595	0.688

Tabela 2. Exemplo de análise da concordância em um texto do *corpus*.

¹ EDU é a menor unidade de texto que pode ser considerada para análise discursiva. Essas unidades representam segmentos básicos do discurso, como frases ou orações independentes, que contribuem para a estrutura retórica do texto.

Observa-se que a análise *silver* resulta consistentemente em valores de concordância mais altos de *Krippendorff Alpha* em relação à análise *gold*, o que é um reflexo da sua natureza mais flexível da proposta. Além disso, o Experimento IV apresenta os resultados mais elevados, o que sugere algum consenso entre os anotadores.

4. Considerações finais

O processo de avaliação da concordância entre anotadores é um aspecto crítico para garantir a qualidade e a consistência dos dados utilizados no treinamento de modelos de AM. Neste trabalho, optamos por um método diferenciado de análise da concordância, propondo duas abordagens distintas: uma análise mais restrita (*gold*) e outra mais flexível (*silver*). Ainda, nessa pesquisa, procuramos prever quatro cenários distintos, em que fosse possível observar se o perfil de anotadores (com mais ou menos experiência com anotação de corpus) poderia influenciar a concordância.

Os resultados apresentados nas Tabelas 1 e 2 demonstram que a abordagem *silver*, devido à sua flexibilidade, tende a produzir valores de concordância mais elevados em comparação à abordagem *gold*, uma vez que considera uma janela de concordância, permitindo uma avaliação mais inclusiva das anotações. Tal aspecto é especialmente útil em cenários em que pode haver pequenas variações nas escolhas de *tokens*, não comprometendo a qualidade da anotação. A adoção dessas duas abordagens permitirá uma compreensão mais abrangente do alinhamento entre os anotadores, oferecendo uma visão tanto da precisão estrita quanto de uma concordância mais ampla e contextual.

A automatização do processo de análise de concordância, como implementado neste trabalho, representa um avanço significativo, permitindo uma avaliação mais rápida e objetiva da qualidade das anotações. O uso das bibliotecas NLTK e lxml para manipulação e análise dos dados mostrou-se extremamente eficaz, o que possibilitou a criação de *pipelines* de processamento que podem ser reutilizados em outros contextos.

Por fim, as metodologias e ferramentas desenvolvidas neste trabalho tem potencial para servir de base para futuras pesquisas. Dessa maneira, a possibilidade de automatizar a análise da concordância acelera o processo de validação das anotações e garante maior objetividade e confiança dos resultados. Tais aspectos são de suma importância em um processo de anotação de um *corpus* e, conseqüentemente, o uso dos dados no treinamento de modelos de AM.

Em trabalhos futuros, caberá um aprofundamento nas análises e métodos propostos inicialmente nesta pesquisa. Destaca-se o fato de haver, como demonstrado, impacto no relaxamento da análise por meio do método *silver*, quando comparado ao método *gold*. Ainda, as versões unificadas dos textos anotados poderão ser analisadas no desempenho da função de adjudicator, ao passo que, ao invés de eliminar os trechos em discordância, o método *silver* pode ser utilizado para validar esses mesmos trechos.

Agradecimentos

Este trabalho foi realizado no âmbito do Centro de Inteligência Artificial da Universidade de São Paulo (C4AI -<http://c4ai.inova.usp.br/>), com o apoio da Fundação de Amparo à Pesquisa do Estado de São Paulo (processo FAPESP #2019/07665-4) e da IBM. Este projeto também foi apoiado pelo Ministério da Ciência, Tecnologia e Inovações, com recursos 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. Além disso agradecemos ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) pelo financiamento e suporte.

Referências

- Bird, S., Klein, E., Loper, E. (2008). *NLTK documentation*. Online. Disponível em: <https://www.nltk.org/>
- Cardoso, P. C., Maziero, E. G., Jorge, M. L. C., Seno, E. M., Di Felippo, A., Rino, L. H. M., ... & Pardo, T. A. (2011). CSTnews-a discourse-annotated corpus for single and multi-document summarization of news texts in Brazilian Portuguese. In *Proceedings of the 3rd RST Brazilian Meeting* (pp. 88-105). Cuiabá/MT.
- Cardoso, P.C.F., Souza, J.W.C., Rodrigues, R. Dantas, E., Cruz, G.S.B., Bárbara, L. de J. S., Gama, N. S., Almeida, T. J. A. Pereira, M.A. 2024. A Linguagem em foco: Anotação de Sinalizadores Discursivos em *corpus* jornalístico. In *Anais do XV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*. Porto Alegre: SBC.
- Dantas, E., Bárbara, L.J.S., Pereira, M.A., Gama, N.S., Almeida, T.J.A., Souza, J.W.C., Cardoso, P.C.F., Rodrigues, R. (2024). *Manual de anotação de sinalizadores discursivos em textos jornalísticos*. São Carlos: Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo. Disponível em <https://repositorio.usp.br/item/003207370>
- Hovy, E., Lavid, J. (2010). Towards a Science of Corpus Annotation: A New Methodological Challenge for Corpus Linguistics. *International Journal of Translation*, 22, p.13-36.
- Krippendorff, K. (2011). *Computing Krippendorff's Alpha-Reliability*. Departmental paper, Annenberg School for Communication, University of Pennsylvania.
- Mann, W.C., Thompson, S. A. (1987). *Rhetorical structure theory: Description and construction of text structures*. In: Natural language generation: New results in artificial intelligence, psychology and linguistics. Dordrecht: Springer Netherlands. p. 85-95.
- McHugh, M.L. (2012). *Interrater reliability: The kappa statistic*. *Biochemia Medica*, 22(3), 276–282
- Passonneau R. (2006) Measuring agreement on set-valued items (MASI) for semantic and pragmatic annotation. In: *Proceedings of the international conference on language resources and evaluation* (LREC). Genoa/Italia: European Language Resources Association. p. 831-836.
- Zeldes, A. (2016) rstWeb-a browser-based annotation interface for Rhetorical Structure Theory and discourse relations. In: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*. p. 1-5.

Mineração de Emoções Multirrotulo Em Textos Curtos

Ramon N. Mendes, Syanne. K. M. Tavares, Luiz Nicollas M. Campos, Fabíola P. O. Araújo

¹Instituto de Ciências Exatas e Naturais – Universidade Federal do Pará (UFPA)
CEP 66.075-110 – Pará – PA – Brazil

{ramon.mendes, syanne.tavares, luiz.campos}@icen.ufpa.br, fpoliveira@ufpa.br

Resumo. *Este artigo apresenta os resultados parciais da investigação de uma abordagem para o reconhecimento de múltiplas emoções em textos curtos em português brasileiro. Para isso, propõe-se a construção de um corpus com tweets coletados por Web Scraper e rotulados com base na teoria da roda de emoções de Plutchik. Além disso, são apresentadas as etapas realizadas no pré-processamento do corpus, no desenvolvimento e na análise comparativa entre os modelos SVM e BERT na detecção de emoções em textos gerados por um LLM. As avaliações demonstraram que o modelo SVM obteve quase 20% a mais de precisão do que o BERT.*

Abstract. *This article presents the partial results of the investigation of an approach for recognizing multiple emotions expressed in short texts in Brazilian Portuguese. To this end, we propose the construction of a corpus with tweets collected by Web Scraper and labeled based on Plutchik's theory of the wheel of emotions. Furthermore, the steps taken in pre-processing the corpus, development and comparative analysis between the SVM and BERT models in detecting emotions in texts generated by an LLM are presented. Evaluations demonstrated that the SVM model achieved almost 20% more accuracy than BERT.*

1. Introdução

A mineração de emoções é o campo de estudo que analisa o aspecto emocional expresso sobre uma diversidade de entidades e atividades humanas [Liu 2012]. Sendo as emoções fatores implícitos na tomada de decisão de qualquer indivíduo, pesquisas procuram entender a proximidade entre decisões cibernéticas e humanas [Pires 2023] e um estudo que vem ganhando destaque nesse contexto é a IA emocional, onde máquinas são treinadas para sentir, detectar, interpretar e analisar emoções. [Mantello et al. 2023].

No processo exploratório da pesquisa, observou-se que a grande quantidade das publicações realizadas na área se limitam em analisar sentimentos classificando-os quando à sua polaridade, podendo ser positiva, negativa ou neutra [Cardozo and Freitas 2021, Paes et al. 2022, Silva and Faria 2023, Seno et al. 2023]. Todavia, sabe-se que as emoções se comportam de forma bem mais complexa. Além disso, apesar da literatura científica ter visto uma proliferação de estudos sobre detecção de emoções em língua inglesa, a quantidade de trabalhos que exploram a temática em textos em português brasileiro (Pt-BR) é ainda escassa [de Oliveira and de Melo 2021, Pereira 2021] e a mineração multilíngue ainda enfrenta desafios em alcançar altas precisões e confiabilidade [Santos et al. 2014, Moreira et al. 2024].

Outra abordagem possível adotada por pesquisas utiliza recursos da aprendizagem profunda, como [Hammes and Freitas 2021], que utilizaram os modelos BERTimbal-base

e BERTimbal-large para classificar 27 emoções de sentenças do *dataset multi-label* Go-Emotions traduzidos para o português. Nele, os autores comentam sobre a indisponibilidade de *datasets* anotados para tarefa de classificação de emoções em português e da perda de qualidade do *dataset* devida a tradução, que pode levar a perda do sentido expresso pelas sentenças e interferência na classificação correta das emoções.

Nessa perspectiva, este trabalho propõe a construção de uma base de dados contendo *tweets* em português rotulados de acordo com suas múltiplas classes de emoções, o treinamento e comparação de modelos para identificar essas emoções no texto. Para alcançar esse objetivo, realizou-se a coleta automática de um grande volume de informações da rede social X, empregando técnicas de processamento de linguagem natural (PLN) e aprendizado de máquina. Além disso, foi realizada uma análise comparativa entre os modelos de máquinas de vetores de suporte (*Support Vector Machine*) e o modelo de aprendizagem profunda BERT (*Bidirectional Encoder Representations from Transformers*) na classificação emocional dos textos da base de dados e em dados não vistos gerados com auxílio de um *Large Language Model* (LLM).

A organização do artigo ocorre da seguinte maneira: na Seção 2, descreve-se a metodologia, incluindo a preparação do corpus, a etapa de pré-processamento, o modelo treinado para a classificação e as métricas de avaliação usadas; na Seção 3, apresenta-se os resultados parciais da análise realizada e discute-se suas implicações; Por fim, na Seção 4, as considerações finais desta pesquisa são apresentadas.

2. Metodologia

No presente trabalho, as emoções foram identificadas de acordo com o modelo do psicólogo norte-americano Robert Plutchik, elaborado em 1980. A roda das emoções é um recurso gráfico em formato de flor de oito pétalas¹. Ela se diferencia das demais teorias ao propor relações de tipologia, antagonismo e intensidade entre as emoções, além de identificá-las. Quanto à tipologia, podem ser 8 emoções primárias (alegria, confiança, medo, surpresa, tristeza, nojo, raiva e antecipação), ou emoções secundárias, que ficam entre as pétalas, geradas pela combinação das emoções primárias adjacentes.

A preparação do corpus é essencial para o desenvolvimento do modelo de classificação de emoções. O conjunto de dados, com 12.160 publicações da rede social X, foi obtido via *Web Scraper* em Python, utilizando sinônimos de emoções primárias e secundárias. As emoções secundárias caracterizaram o conjunto como multirrótulo.

A rotulação de cada *tweet* teve como base as palavras sinônimas utilizadas no *Web Scraper*. Por exemplo, o texto "Do nada bate um desânimo!" foi coletado por meio do sinônimo "Desânimo" e foi classificado como Tristeza. Já "Não tenho provas, tenho convicção" foi coletado por meio de "Convicção" e foi classificado como confiança.

O pré-processamento do texto é essencial na análise de emoções, pois prepara e limpa os dados para um processamento mais eficiente. Primeiramente, são removidos caracteres especiais e acentos, especialmente em idiomas como o português. Depois, ocorre a tokenização, onde o texto é dividido em palavras (*tokens*), descartando caracteres e números restantes. Em seguida, retiram-se as *stopwords*, palavras comuns que pouco

¹https://github.com/MiningEmotion/EmotionsMiningPTBR/blob/main/imagens/Roda_Das_Emocoes.png

contribuem para o contexto. Por fim, utiliza-se o *stemming* para reduzir palavras ao seu radical, normalizando variações verbais e nominais.

Após a limpeza dos dados textuais descrita acima, aplicou-se a técnica de vetorização com a técnica *TF-IDF* (*Term frequency - inverse document frequency*), que considera a frequência e importância das palavras para que se possa aplicar um classificador e determinar a predição final do texto [LANDIM and TRESSO 2023].

3. Modelos de Aprendizagem de Máquina

Foram utilizados dois modelos de aprendizado de máquina para as fases de treino e teste, ambos com as mesmas métricas para uma melhor comparação. Os modelos empregados foram o *Support Vector Machine* (SVM) e o *Bidirectional Encoder Representations for Transformers* (BERT).

O modelo SVM linear foi utilizado em conjunto com o *MultiOutputClassifier*² que consiste em treinar um classificador separado por cada rótulo, adaptando o classificador base que não suporta nativamente uma saída com mais de um rótulo, como o SVM, para um modelo de classificação multirrótulo.

O BERT é baseado na arquitetura *transformer* [Devlin et al. 2018], ao contrário dos modelos direcionais, que lêem a entrada de texto sequencialmente, da esquerda para a direita ou da direita para a esquerda, o codificador *transformer* lê toda a sequência de palavras de uma vez, por isso é considerado bidirecional, também utilizado na sua versão multirrótulo e treinado com 4 épocas .

Ambos os modelos foram treinados com o conjunto de dados coletados e testados com dados gerados por um LLM, onde foram geradas 10 frases de cada emoção secundária para testar o modelo treinado, categorizando um texto com múltiplas classes simultaneamente. O *prompt* utilizado na criação das frases foi "Gere 10 frases com as emoções primárias : 'alegria' e 'confiança' e com a emoção secundária 'amor' que imite a mesma linguagem informal de tweets".

4. Resultados e Discussões

Visando avaliar os resultados e o desempenho dos modelos, foram adotadas as seguintes métricas para a análise da classificação multirrótulo: acurácia, precisão, revocação e F1. A seleção das métricas se deu considerando a especificidade de cada uma delas.

De acordo com a Figura 1, o modelo SVM demonstrou uma precisão significativa de 85,18%, indicando uma boa capacidade de classificação. No entanto, o BERT ao ser confrontado com as mesmas frases não vistas geradas pelo LLM, a acurácia caiu drasticamente para 66,67%, sinalizando dificuldades na classificação desses dados. Além disso, a acurácia, revocação e métrica F1 também diminuíram consideravelmente para as frases geradas, indicando uma tendência do segundo modelo em cometer mais erros de classificação e em perder instâncias de emoções.

A diminuição nas métricas pode ser atribuída à natureza dos modelos. O SVM tenta encontrar um hiperplano com a maior margem de separação, assegurando que os dados de cada classe sejam classificados corretamente com maior probabilidade. Já o BERT

²<https://scikit-learn.org/stable/modules/generated/sklearn.multioutput.MultiOutputClassifier.html>

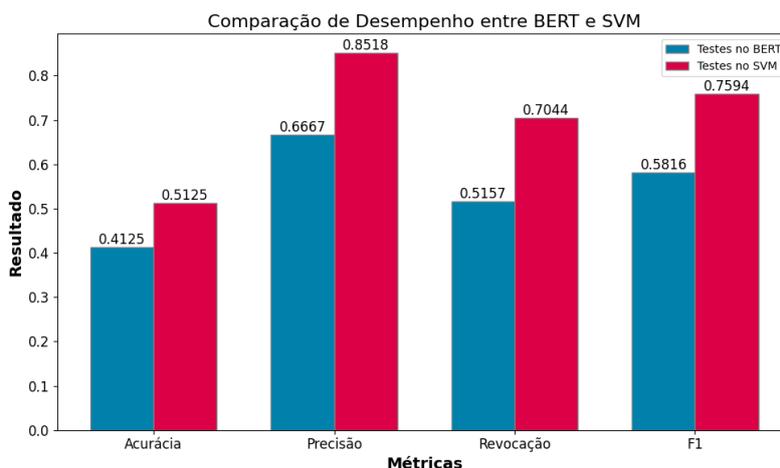


Figura 1. Comparação de Desempenho entre BERT e SVM. (Fonte: Autores).

utiliza o treinamento bidirecional do *transformer*, permitindo que o modelo aprenda o contexto de uma palavra com base em seu entorno completo.

No modelo SVM, quanto menor o texto, menor será o processamento exigido pela máquina. Em contrapartida, o BERT, sendo um modelo de linguagem mais robusto, necessita de textos mais extensos para utilizar todo o seu potencial. Conforme demonstrado por [Devlin et al. 2018], ao empregar características semânticas dos revisores junto com textos curtos, observou-se uma melhoria na acurácia da classificação do modelo BERT, enquanto o SVM não apresentou mudanças significativas nos valores das métricas.

5. Considerações Finais e Direções Futuras

Este trabalho contribuiu na investigação referente à classificação multirrotulo de emoções em textos curtos com a disponibilização de códigos, a base de dados *tweetEmotionsPTBR* e as frases geradas por um LLM, já classificados em português brasileiro, no repositório *EmotionMiningPTBR*³ do Github.

Através dos resultados, o SVM obteve maior desempenho em relação ao BERT na maioria das métricas devido à natureza do modelo. Todavia, as métricas não desempenharam conforme esperado pelo modo como os modelos foram treinados, exclusivamente com conteúdo da rede social X, carecendo de uma quantidade significativa de exemplos que contenham uma linguagem mais elaborada e nuances figurativas. Sendo assim, foram realizados testes iniciais com conjuntos de dados maiores e balanceados, treinando os modelos com as publicações de redes sociais e frases geradas por LLM. Notou-se uma melhora significativa nas métricas do modelo BERT, superando o SVM.

Como direções futuras, é necessário a construção de um conjunto de dados mais diversificado linguisticamente, proveniente de uma variedade de fontes. Isso ajuda a evitar que os modelos fiquem excessivamente adaptados a um único estilo de linguagem. Além disso, seria importante a elaboração de um *Gold-Standard Corpus* com a revisão dos rótulos feita por especialistas na teoria da roda das emoções.

³<https://github.com/MiningEmotion/EmotionsMiningPTBR>

Referências

- Cardozo, L. and Freitas, L. (2021). Análise de sentimentos: Avaliando o desempenho de pré-processamento e de algoritmos de aprendizagem de máquina sobre o dataset tweetsentbr. In *Anais do X Brazilian Workshop on Social Network Analysis and Mining*, pages 169–174, Porto Alegre, RS, Brasil. SBC.
- de Oliveira, M. and de Melo, T. (2021). An empirical study of text features for identifying subjective sentences in portuguese. In Britto, A. and Valdivia Delgado, K., editors, *Intelligent Systems*, pages 374–388, Cham. Springer International Publishing.
- Devlin, J., Chang, M., Lee, K., and Toutanova, K. (2018). BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.
- Hammes, L. and Freitas, L. (2021). Utilizando bertimbau para a classificação de emoções em português. In *Anais do XIII Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 56–63, Porto Alegre, RS, Brasil. SBC.
- LANDIM, G. P. P. and TRESSO, G. J. (2023). Identificação de sentimentos em textos utilizando o modelo term frequency-inverse document frequency.
- Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies*, 5:1–167.
- Mantello, P., Ho, M.-T., Nguyen, M.-H., and Vuong, Q.-H. (2023). Machines that feel: behavioral determinants of attitude towards affect recognition technology—upgrading technology acceptance theory with the mindsponge model. *Humanities and Social Sciences Communications*, 10(1):1–16.
- Moreira, B. G., Pfitscher, R. J., Camargo, L. C., and Garcia, T. R. (2024). Análise de sentimentos e emoções com o uso da ferramenta orange datamining: uma avaliação a partir da tradução dos textos do português. *Anais do Computer on the Beach*, 15:352–354.
- Paes, V., Araújo, D., Brito, K., and Andrade, E. (2022). Análise de sentimento em tweets relacionados ao desmatamento da floresta amazônica. In *Anais do XI Brazilian Workshop on Social Network Analysis and Mining*, pages 61–72, Porto Alegre, RS, Brasil. SBC.
- Pereira, D. A. (2021). A survey of sentiment analysis in the portuguese language. *Artif. Intell. Rev.*, 54(2):1087–1115.
- Pires, S. F. S. (2023). Inteligência artificial e emoções: Quão próximas estão as decisões cibernéticas das humanas? *A sociedade do conhecimento e suas tecnologias: estudos em Ciências Exatas e Engenharias: Volume 10*.
- Santos, A., Becker, K., and Moreira, V. (2014). Um estudo de caso de mineração de emoções em textos multilíngues. In *Anais do III Brazilian Workshop on Social Network Analysis and Mining*, pages 140–151, Porto Alegre, RS, Brasil. SBC.
- Seno, E., Anno, F., Lazarini, L., and Caseli, H. (2023). Classificação de polaridade orientada aos alvos de opinião em comentários sobre debate político em português. In *Anais do XIV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 84–93, Porto Alegre, RS, Brasil. SBC.

Silva, S. and Faria, E. (2023). Análise de sentimentos expressos no twitter em relação aos candidatos da eleição presidencial de 2022. In *Anais do XII Brazilian Workshop on Social Network Analysis and Mining*, pages 79–90, Porto Alegre, RS, Brasil. SBC.

Classificação automática de textos de *User-Generated Content* utilizando Aprendizagem de Máquina Supervisionado

Iolanda Victoria Morais Ramos ¹, Jackson Wilke da Cruz Souza ^{1,2}

¹Instituto de Ciência, Tecnologia e Inovação - Universidade Federal da Bahia (UFBA), Camaçari/BA - Brasil

²Programa de Pós-Graduação em Língua e Cultura - Universidade Federal da Bahia (UFBA), Salvador/BA - Brasil

iolanda.ramos@ufba.br, jackcruzsouza@gmail.com

Abstract. *This study aims to develop an automatic text classifier for User-Generated Content from the DANTE-Stocks corpus. The classification algorithm was trained in a supervised manner, using labels provided by human annotators and subsequently associated with various vectorization methods. In the end, a classifier was generated that performs very close to human-level performance in identifying the three proposed classes, namely: (i) well-structured, (ii) moderately structured, and (iii) poorly structured.*

Resumo: *Este estudo visa a construção de um classificador automático de textos User-Generated Content do corpus DANTE-Stocks. O algoritmo de classificação foi treinado de forma supervisionada, utilizando rótulos propostos por anotadores humanos e, posteriormente, associado a diferentes métodos de vetorização. Ao final, gerou-se um classificador que performa bastante próximo ao desempenho humano, ao identificar três classes propostas de organização dos tweets, a saber: (i) bem, (ii) mediantemente e (iii) mal estruturado.*

1. Introdução

As redes sociais têm desempenhado um papel crucial para a produção, circulação e recepção de conteúdos de interesses para a sociedade. Com a expansão das redes sociais, os usuários assumiram um papel cada vez mais ativo como geradores de conteúdo. Os conteúdos gerados por usuários (em inglês, *User-Generated Content* (UGC)), segundo Wyrwoll (2014), podem significar uma grande contribuição para o desenvolvimento e progresso intelectual da sociedade.

Para a área de Processamento de Línguas Naturais (PLN), UGC apresenta desafios únicos de processamento dadas suas características ligadas à linguagem e ao modo de circulação de mensagens. Nesse sentido, o conteúdo gerado pode não seguir padrões linguísticos e estruturais ligados à norma culta da língua, apresentando grande diversidade nessas questões.

Para lidar com esses desafios, é necessário um conjunto de técnicas e recursos em PLN, sobretudo para o Português do Brasil (PB), língua ainda em fase de desenvolvimento de recursos para o processamento de textos de UGC. Técnicas como, por exemplo, classificação e agrupamentos desses textos são de grande importância para aprimorar a *qualidade* de identificação de padrões e fenômenos linguísticos, e as *dimensionalidades* quanto à gerenciabilidade da performance dos modelos, facilitando, posteriormente, análises e modelagem linguístico-computacionais.

Partindo da classificação prévia de Pereira e Souza (2024), construímos um classificador automático para *tweets*/postagens do *corpus* DANTEStock [Di Felippo *et al.* 2021]. Foram testadas duas abordagens de vetorização dos dados linguísticos para a construção dos algoritmos de Aprendizado de Máquina (AM) supervisionado a partir do paradigma teórico conexionista/neural [Monard e Baranauskas 2003], o qual busca simular o processamento de informações inspirado no modelo biológico do sistema nervoso. Além disso, as classificações foram submetidas a avaliações *quantitativas* (a partir de métricas clássicas de AM) e *qualitativas* (avaliação humana). Como resultado, foi promovida a classificação do *corpus* em função da estrutura de cada *tweet*/postagem.

Destaca-se que o emprego de diferentes técnicas de avaliação e desenvolvimento de ferramentas e recursos em PLN devem ser compreendidos como uma ponte fundamental entre linguística e computação. Rodrigues, Souza e Santos (2022) destacam que essa interação é “de mão dupla”. Isso significa que, por um lado, a linguística é essencial para desenvolver e melhorar os recursos que as máquinas usam para interpretar a linguagem. Por outro lado, as ferramentas computacionais também podem ajudar a refinar e validar os conhecimentos linguísticos criados pelos humanos.

Para tanto, este artigo está organizado em cinco seções, além desta Introdução. Na Seção 2, apresentamos a metodologia utilizada nesta pesquisa. Na Seção 3, destacamos os resultados no processo de treinamento do modelo supervisionado desenvolvido para a tarefa de classificação de textos de UGC. Por fim, na Seção 4, tecemos considerações finais e indicações de trabalhos futuros.

2. Metodologia

Neste trabalho, propusemos a criação de um modelo de AM supervisionado para classificação do *corpus* DANTEStock [Di Felippo *et al.* 2021]. Tal *corpus* é constituído por *tweets* ligados ao cenário de ações da bolsa de valores do Brasil. O *corpus* é composto por 4,518 *tweets* e seus identificadores únicos, que foram compilados a partir da coleta automática de postagens do X/Twitter, em 2014.

Para a criação e o treinamento de um classificador neste trabalho, foram testados algoritmos de diferentes paradigmas, sendo o algoritmo *Multilayer classifier* - MLP [Haykin 1994] o escolhido por apresentar melhor desempenho em termos de métricas quantitativas e avaliação qualitativa. Destaca-se que o ambiente de desenvolvimento foi o Colaboratory do Google. Ademais, as bibliotecas utilizadas em Python foram extraídas do *scikit-learn* [Kramer 2016].

Para a tarefa de treinamento do modelo, uma amostra menor do *corpus* contendo 180 *tweets* foi rotulada por três anotadores, como apontado por Pereira e Souza (2024). Neste trabalho, os autores propuseram três classes considerando a organização sintática, semântica e estrutural das sentenças, a saber: *bem estruturado*, com 81 exemplares, *mediamente estruturado*, com 59, e *mal estruturado*, com 39. Os algoritmos foram treinados observando o texto dos *tweets* e tendo como alvo de predição as classes propostas pelos anotadores, como exemplificado em (1), retirado de Pereira e Souza (2024).

(1)

a) *Bem estruturado*: Ano passado eu falei que até o final de 2104 #PETR4 estaria abaixo de R\$10,00 mas acho que errei, não vai demorar tanto.

b) *Mediamente estruturado*: vai, oibr4. um troux... ops... investidor precisa pagar as minhas férias

c) *Mal estruturado*: <,Alexander Cruz3 *-*

O *tweet* (1a) foi classificado como *bem estruturado*, pois sua estrutura não prejudica em nada a compreensão do conteúdo, mesmo sem um contexto informacional maior. Já o *tweet* (1b) foi classificado como *mediamente estruturado*, já que possui compreensão limitada dada a sua (des)organização sintática e semântica. Por fim, o *tweet* (1c), classificado como *mal estruturado*, tem uma baixa compreensão em aspecto semântico e falta estruturação lógica que permita a compreensão da mensagem.

Para lidar com o desbalanceamento dos dados, o modelo foi desenvolvido considerando as técnicas de validação cruzada [Netto e Maciel 2021] para melhorar o aprendizado dos critérios de cada classe e a vetorização das instâncias com base no modelo pré-treinado de Bertimbau [Souza e Nogueira 2020].

Após a etapa de treinamento do modelo, passamos para a etapa de análise quanti e qualitativa dos resultados de desempenho de cada modelo. Nessa etapa, foram avaliadas quantitativamente as métricas de desempenho [Netto e Maciel 2021], com base em Precisão (P), Revocação (R), Medida-F (MF) e Acurácia (A). Quanto à avaliação qualitativa, selecionamos de forma aleatória um exemplo de *tweet*/postagem e analisamos se a classificação proposta pelo modelo fazia sentido se comparada com a avaliação humana da estrutura sentencial do exemplo.

3. Resultados e discussão

De forma preliminar, cada modelo classificador foi avaliado de acordo com suas métricas utilizadas neste trabalho (P, R, MF e A). O classificador foi treinado utilizando duas diferentes técnicas de vetorização, a saber *Term Frequency - Inverse Data Frequency - TF-IDF* [Moreira 2024] e *Bidirectional Encoder Representations from Transformers (BERT)*, mais especificamente usando a variação treinada para o PB, o *BERTimbau*. Em ambos os casos, o modelo MLP se mostrou mais adequado. A Tabela 1 resume o resultado dos desempenhos obtidos ao classificar nossa amostra.

Classes	Classificador / Medidas							
	MLP - TF-IDF				MLP - BERT			
	P	R	MF	A	P	R	MF	A
Bem estruturado	0.58	0.72	0.64	0.57	0.77	0.80	0.78	0.74
Mediamente estruturado	0.47	0.35	0.40	0.57	0.68	0.65	0.67	0.74
Mal estruturado	0.75	0.67	0.71	0.57	0.78	0.78	0.78	0.74

Tabela 1. Métricas obtidas de cada modelo em etapa inicial

O desempenho do modelo MLP pode, possivelmente, ser explicado pelo próprio paradigma conexionista que o modelo possui, sendo capaz de absorver melhor as nuances mais complexas de representações propostas na etapa de aprendizado baseada nos rótulos fornecidos. Outro aspecto importante a ser considerado é a melhora significativa de desempenho do modelo quando associado ao método de vetorização utilizando BERT. Em termos de acurácia, os modelos saíram de 57% para 74%; já em termos de MF, a classe que mais bem foi beneficiada com a abordagem foi a de textos *mediamente estruturados*, saindo de 40% para 67%. Essa melhora pode possivelmente ser explicada

pela capacidade do BERT de entender o contexto das palavras, o que é crucial para a classificação de *tweets*, em que o contexto pode alterar fortemente o significado. Além disso, o BERTimbau é um modelo não apenas treinado para o PB, mas também treinado com conteúdo proveniente de redes sociais. Isso faz com que o classificador lide melhor com as nuances semânticas captadas pela vetorização.

Além disso, submetemos o classificador a novos exemplos de *tweets*/postagens retirados para avaliarmos a capacidade de generalização, como exemplificado em (2).

(2)

a) *Bem estruturado*: Vamos ver se mesmo em dia fraco, ganha alguma força ...RT @Live_Trade: #ecor3 fechando 15' acima 12,53 já fica interessante”

b) *Mediamente estruturado*: que linda era esa mina chabonnn

c) *Mal estruturado*: @victoriabril_ forra contesta

Nos exemplos, a categoria atribuída pelo modelo MLP a (2a) foi *bem estruturada*: a estrutura do *tweet* não prejudica sua compreensão, embora apresente desvios de pontuação, por exemplo. Por outro lado, em (2b) o modelo classificou a instância como *mal estruturada*, o que difere da avaliação humana, que rotulou a mesma instância como *mediamente estruturada*. Apesar dos desvios ortográficos presentes, é possível identificar uma estrutura mínima na mensagem, a qual poderia ser mais bem compreendida se considerada dentro de um contexto adequado. Por fim, (2c) foi classificado como *mal estruturado*: a estrutura do *tweet* não permite nenhuma compreensão acerca da mensagem.

4. Considerações finais

Os objetivos deste trabalho, que incluíam a construção de um classificador automático para *tweets* e postagens do *corpus* DANTEStock, a aplicação de técnicas de AM e a avaliação das classificações, foram devidamente alcançados. Por se tratar de um estudo preliminar, optou-se por trabalhar com uma amostra reduzida do *corpus* original, em virtude do alto custo de recurso humano para a classificação manual dos dados, essencial para garantir a precisão e confiabilidade das informações. Apesar desta questão, os resultados quantitativos e qualitativos demonstraram a capacidade do classificador de lidar com a categorização de textos de UGC. Além disso, o agrupamento do *corpus* com base na estrutura sentencial dos textos tornará possível anotações linguísticas e/ou a identificação de padrões importantes sobre a diversidade e complexidade dos textos.

Para a tarefa de classificação de textos de UGC, o uso de AM supervisionado permite que o modelo aprenda de maneira aproximada ao desempenho humano. Trata-se do levantamento de características importantes em função do rótulo-alvo, melhorando a precisão com que o modelo performa e generaliza em exemplos semelhantes, *a posteriori*.

Destaca-se que esta pesquisa contribui de maneira significativa para o projeto POeTiSA: PORTuguese processing - Towards Syntactic Analysis and parsing, que visa desenvolver ferramentas e aplicações linguístico-computacionais para o PB. A integração da análise de UGC em diferentes teorias linguístico-computacionais, pode auxiliar, por exemplo, na identificação de fenômenos ainda não descritos no PB. Para trabalhos futuros, serão explorados outros algoritmos de AM, além de desenvolver estratégias para melhorar o balanceamento dos textos, garantindo a preservação da naturalidade dos dados durante o processo de ajuste e modelagem.

Agradecimentos

Este trabalho foi realizado no âmbito do Centro de Inteligência Artificial da Universidade de São Paulo (C4AI -<http://c4ai.inova.usp.br/>), com o apoio da Fundação de Amparo à Pesquisa do Estado de São Paulo (processo FAPESP #2019/07665-4) e da IBM. Este projeto também foi apoiado pelo Ministério da Ciência, Tecnologia e Inovações, com recursos 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. Este projeto também foi apoiado pela Universidade Federal da Bahia, através do programa de bolsas de iniciação científica - ações afirmativas (PIBIC-AF) 2023/2024.

Referências

- Breiman, L. (2001). Random Forests. *Machine Learning*, 45, pp. 5–32. DOI: <https://doi.org/10.1023/A:1010933404324>
- Di Felippo, A. *et al.* (2021). “Descrição Preliminar do Corpus DANTEStocks: Diretrizes de Segmentação para Anotação segundo Universal Dependencies”. In: *Anais do XIII Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*. Porto Alegre, Brasil: SBC, p. 335-343. DOI: <https://doi.org/10.5753/stil.2021.17813>
- Haykin, S. (1994). *Neural Networks: A Comprehensive Foundation*. Prentice Hall PTR.
- Kramer, O., e Kramer, O. (2016) Scikit-learn. *Machine learning for evolution strategies*, p. 45-53
- Mann, W. C., e Thompson, S. A. (1988). Rhetorical Structure Theory: Toward a Functional Theory of Text Organization. *Text-interdisciplinary Journal for the Study of Discourse*, 8(3), pp. 243–281. Disponível em: https://www.sfu.ca/rst/05bibliographies/bibs/Mann_Thompson_1988.pdf
- Mikolov, T. *et al.* (2013). Efficient Estimation of Word Representations in Vector Space. *Preprint*. Disponível em: <http://arXiv:1301.3781>
- Monard, M. C., E Baranauskas, J. A. (2003). Conceitos sobre Aprendizado de Máquina. *Sistemas Inteligentes: Fundamentos e Aplicações*, 1(1), p. 1.
- MOREIRA, V. P. (2024). Recuperação de Informação. In: NUNES, M. G. (Org.), *Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português*, 2. ed. [s.l.]: BPLN. Disponível em: <https://brasileiraspln.com/livro-pln/2a-edicao/parte-aplicacoes/cap-ir/cap-ir.html>
- NETTO, A., e MACIEL, F. (2021). *Python para Data Science e Machine Learning Descomplicado*. Rio de Janeiro: Editora Alta Books, 397p.
- Pereira, M.A., Souza, J.W.C. (2024). Subsídios Linguísticos para classificação automática de textos de User-Generated Content. In *Anais do XV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*. Porto Alegre: SBC.
- Rodrigues, R., Souza, J. W. C., e Santos, R. L. S. (2022). “Descrição Linguística e Aprendizado de Máquina: Análise de Verbos Locativos do Espanhol”. *Cadernos de Estudos Linguísticos*, Campinas, SP, 64(00), p. e022038. DOI: <https://doi.org/10.20396/cel.v64i00.8666995>

Souza, F., Nogueira, R., E Lotufo, R. (2020). BERTimbau: Pretrained BERT Models for Brazilian Portuguese. In: Cerri, R., and Prati, R. C. (Eds.), *Intelligent Systems. BRACIS 2020. Lecture Notes in Computer Science*, vol. 12319, Springer, Cham. DOI: https://doi.org/10.1007/978-3-030-61377-8_28

Wyrwoll, C. (2014). *User-Generated Content. Social Media*. In C. Wyrwoll (Ed.), *Social Media: Fundamentals, Models, and Ranking of User-Generated Content*. Springer Fachmedien, p 11–45. DOI: https://doi.org/10.1007/978-3-658-06984-1_2

Zhang, H. (2004). The Optimality of Naive Bayes. In: *Proceedings of the Seventeenth International Florida Artificial Intelligence Research Society Conference*. Florida/USA: American Association for Artificial Intelligence. p.1-6.

PLN e Segurança Jurídica

Identificação de divergências jurisprudenciais com Processamento de Linguagem Natural

Marcella Queiroz de Castro¹; Ana Régia Mendonça¹

¹Instituto Federal de Educação, Ciência e Tecnologia de Brasília (IFB)
Brasília, DF – Brasil
marcella.castro@estudante.ifb.edu.br

Abstract. *This paper proposes the use of Natural Language Processing (NLP) and Machine Learning techniques to identify judicial rulings that are divergent to the majority understanding of the subject in Brazilian courts, aiming to enhance judicial security. The methodology used includes data preprocessing of a summary of the rulings, the use of the Word2Vec neural network technique for word embedding, and the analysis of 3.165 court decisions via k-means clustering to identify semantic similarities and divergences. Specific examples of jurisprudential divergences are presented, demonstrating how technology can assist in the uniformity of judicial decisions.*

Resumo. *Este artigo propõe a utilização de técnicas de Processamento de Língua Natural (PLN) e Aprendizado de Máquina para identificar divergências jurisprudenciais nos tribunais brasileiros, visando aumentar a Segurança Jurídica. A metodologia inclui a extração de 3.165 acórdãos do Tribunal de Justiça de Minas Gerais, o pré-processamento dos dados, uso da técnica de Word2Vec para definição de embeddings para cada palavra e análise de acórdãos via clustering para identificar semelhanças textuais e divergências nas decisões judiciais. Exemplos específicos de divergências jurisprudenciais são apresentados, demonstrando como a tecnologia pode auxiliar na uniformização das decisões.*

1. Introdução

Atualmente, no caso brasileiro, espera-se de uma IA (Inteligência Artificial) no judiciário o auxílio aos servidores para superar o enorme acervo de processos que aguardam julgamento, almejando celeridade na tramitação processual. Conforme o relatório Justiça em Números do Conselho Nacional de Justiça (CNJ), em 2023, os 92 tribunais do país possuíam pendentes de julgamento 81,4 milhões de casos [Brasil 2023], um cenário de contingente de processos alarmante, em que se faz essencial o desenvolvimento de ferramentas que não somente assegurem celeridade do processo, mas que garantam a estabilidade, previsibilidade e coerência das decisões proferidas pelos tribunais.

Com o intuito de incrementar a Segurança Jurídica das Cortes brasileiras, propõe-se a utilização de técnicas de PLN e Aprendizado de Máquina para apontar divergências jurisprudenciais. Para tanto, o trabalho realizou a extração de acórdãos do Tribunal de Justiça de Minas Gerais (TJMG) com python¹, o pré-processamento e passagem desses pelo processo de definição de *Embeds* via a técnica de rede neural chamada *Word2Vec* [Mikolov et al. 2013].

¹O código de raspagem e análise dos dados pode ser encontrado neste link: <https://colab.research.google.com/drive/1GLp9jU1qMLKdf8P6OPVo-QkdFcp9KCp1?usp=sharing>.

Uma vez que os textos passaram pela etapa de extração de características, tornou-se possível a construção de ferramenta de análise e classificação de acórdãos em *clusters*, capazes de identificar semelhanças entre as decisões por método de Similaridade Textual Semântica (STS) e agrupá-las de acordo com o grau de similaridade fático e decisório que possuem, apontando possíveis divergências jurisprudenciais no conjunto de decisões em estudo.

2. Conceitos Jurídicos

Um acórdão é todo pronunciamento judicial proferido por um órgão colegiado [Didier 2019] e está, usualmente, contido em um documento de cinco partes, das quais a mais relevante para este estudo é a ementa. A ementa é um resumo da decisão, mostrando quais argumentos foram acolhidos e a decisão tomada pelos magistrados. Estudos analisados [Wilton 2022, Gomes 2021] e a empiria da prática judicial por aplicadores do Direito demonstram que a ementa é suficiente para identificar os termos principais do acórdão, razão pela qual somente essa parte do todo foi selecionada para a análise por PLN. Ademais, são os acórdãos que possuem a tarefa legal de trazer estabilidade ao sistema de precedentes judiciais², demonstrando-se como adequado o estudo de acórdãos para a averiguação da Segurança Jurídica de uma Corte.

3. Análise das Ementas

3.1. Pré-processamento e extração de características

Assim como em trabalhos correlatos [Ciurlino 2021, Polo et al. 2021], as decisões judiciais utilizadas para análise foram obtidas pela extração automatizada a partir de *scripts*, almejando viabilizar a execução em larga escala e para outros temas além do escolhido nesta pesquisa. Para selecionar os dados foi feita pesquisa no site do Tribunal de Justiça de Minas Gerais³ pelas palavras chave “direito do consumidor” e “apelação” nas ementas dos acórdãos. O período de busca escolhido foi de 1º de janeiro de 2021 até 31 de dezembro de 2023, resultando em 3.163 acórdãos extraídos. Uma vez selecionado o *corpus*, é preciso executar a limpeza do texto escolhido. No caso da busca por divergências jurisprudenciais, o contexto jurídico e as práticas dentro de textos desse tipo foram determinantes para definição das etapas de pré-processamento textual.

Para além do pré-processamento e da normalização, os textos passaram por outras etapas prévias que merecem destaque. Palavras comuns do discurso jurídico foram removidas, como “excelência” e “douto”, e, além dos n-gramas existentes dentro do modelo de linguagem adotado, foram adicionados outros específicos do *corpus*, como “responsabilidade_fornecedor”. Inspirado no trabalho correlato de Martins [2018], o qual incluiu formas de pré-processamento como a remoção de textos entre parênteses a partir de análises sobre a forma de redação de documentos jurídicos, este trabalho, de uma análise do *corpus* escolhido, verificou que a prática de escrita dos desembargadores do TJMG inclui colocar entre parênteses ou aspas referências a outros acórdãos citados, a artigos de lei ou a súmulas de outros tribunais, informações estas que também foram retiradas por serem consideradas irrelevantes para a análise computacional.

Quanto à vetorização das palavras, estudos que argumentam pela utilização conjunta de um modelo Word2Vec [Mikolov et al. 2013] com *corpus* de treinamento jurídico

²Art. 926: “Os tribunais devem uniformizar sua jurisprudência e mantê-la estável, íntegra e coerente” [Brasil 2015]

³O site utilizado foi: <https://www5.tjmg.jus.br/jurisprudenciaformEspelhoAcordao.do>

para alcançar o melhor desempenho na análise de similaridade de textos [Xia et al. 2019] e também que destacam como é essencial o treinamento do modelo em uma base jurídica para realização de tarefas específicas de cada tipo de domínio de conhecimento [Gomes 2021] foram norteadores para decidir o tipo de modelo, arquitetura e *corpus* de treinamento.

Após pesquisa sobre os modelos disponíveis para utilização, optou-se pelo modelo LegalNLP [Polo et al. 2021]. Os pesquisadores desse projeto apresentaram um modelo já pré-treinado com textos jurídicos, mais especificamente, com decisões judiciais de Cortes Estaduais, de forma que a ferramenta está alinhada com os textos que serão analisados por similaridade textual neste artigo. Dentre as versões do modelo Word2Vec do projeto, escolheu-se a arquitetura Skip-gram [Mikolov 2013], com *size=200*, *window=15* e *epochs=20*.

3.2. Análise de Similaridade Textual Semântica

Em estudos da área foi difundido o cálculo de similaridade por 'soft cosine' em que fora feita modificação no cálculo da similaridade por cosseno para passar a considerar uma matriz de similaridade das características, a qual acrescenta um peso que aproxima ou afasta os vetores a depender da similaridade de cada palavra do documento [Sidorov 2014].

Essa forma de cálculo e abordagem para o problema da similaridade textual semântica se demonstra útil em cenários onde palavras diferentes podem ter significados semelhantes, como nos textos jurídicos. Por exemplo, "provimento" e "desprovimento" são palavras que ocupam espaços semelhantes dentro da vetorização do modelo pois, ainda que sejam opostas, seguindo a lógica do modelo Word2Vec, costumam vir acompanhadas de outras palavras idênticas. Para combater essa particularidade do contexto jurídico, o cálculo de similaridade por 'soft cosine' se demonstra adequado, pois permite que características relacionadas contribuam para o cálculo de similaridade dos vetores, gerando uma medida de similaridade textual robusta e focada na semântica do texto.

3.3. K-Means

Para lidar com vetores de múltiplas dimensões e focar somente nas mais significativas foi aplicada a técnica de t-SNE (t-Distributed Stochastic Neighbor Embedding) [Maaten 2008], visando permitir uma melhor representação bidimensional do objeto de estudo. A redução de dimensionalidade é etapa considerada relevante em estudos conexos [Wilton 2022, Xia et al. 2019, Martins 2018] e foi seguida neste artigo para garantir uma análise atualizada ao estado da arte da área.

Após a redução de dimensionalidade, o algoritmo do K-Means [Pajankar 2022] necessita que seja definido o número *k* ideal de *clusters*. Para tal, foi utilizado o coeficiente de *Silhouette* [Rousseeuw 1987], que é uma métrica interna que avalia a qualidade dos *clusters* formados pelo algoritmo K-Means, medindo tanto a coesão quanto a separação dos *clusters*. Os parâmetros para a construção da curva do coeficiente de *Silhouette* para os dados deste estudo seguiram as indicações de estudos símiles [Magalhães 2023, Lima 2022], com *n_init=10*, *random_state=0* e *max_iter=300*. O melhor coeficiente de *Silhouette* obtido foi 0.48 para um número de 15 grupamentos e o resultado encontra-se na Figura 1.

4. Divergências Jurisprudenciais

Com base na Figura 1, a investigação foi voltada para os *clusters* com pontos mais dispersos, como os grupos 3 e 11, na busca por divergências jurisprudenciais. Analisando-se manualmente os grupos citados, foram encontrados pontos de retoque na jurisprudência do Tribunal para que esta se torne mais coesa e uniforme.

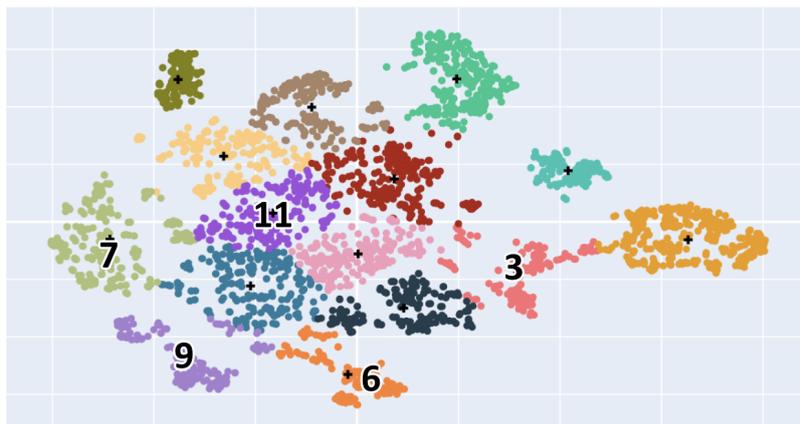


Figura 1. K-Means com 15 clusters realizado sobre as ementas selecionadas e reduzidas por técnica de t-SNE

No cluster 3, comparam-se as ementas proferidas nos processos 5000575.70.2020.8.13.0393 (ementa nº 719 da extração) e 5001067.08.2022.8.13.0453 (ementa nº 496 da extração). Em ambos os casos, trata-se de um contrato de crédito firmado sem o conhecimento da pessoa e que gerou descontos indevidos em seus benefícios previdenciários. Para o primeiro acórdão o desconto indevido só é capaz de gerar dano moral quando comprometer parcela significativa dos proventos. No outro acórdão, no entanto, foi proferido entendimento de que o desconto indevido gera dano moral por decorrência do próprio fato, independente do montante comprometido. De igual maneira, no cluster 11, encontram-se acórdãos que suscitam questionamentos sobre a uniformidade dos julgamentos. Em um caso, é determinada a repetição em dobro de valores que foram cobrados indevidamente a um consumidor, considerando somente a comprovação de que os valores não eram corretos de serem cobrados (Apelação Cível 5016736.60.2019.8.13.0145, ementa nº 2317 da extração). Noutro acórdão, do mesmo cluster, exige-se do consumidor que comprove a má-fé da empresa que realizou a cobrança indevida para ser ressarcido em dobro (Apelação Cível 6787163.48.2009.8.13.0024, ementa nº 3055 da extração).

Sem o objetivo de exaurir as divergências jurisprudenciais possíveis de serem encontradas nessa amostra de ementas, mostra-se interessante para a análise que, de uma verificação manual feita por pessoa formada na área jurídica, sem automatização de qualquer tipo, foi possível encontrar com relativa facilidade precedentes em que situações semelhantes se depa-ram com decisões de teor significativamente distinto. A incursão no conjunto de dados, embora manual, foi orientada pelas etapas prévias de PLN e Aprendizagem de Máquina, demonstrando a capacidade da tecnologia de auxiliar na identificação de uma quebra do princípio da Segurança Jurídica.

5. Conclusões

Inspirada na problemática de falta de Segurança Jurídica, o presente estudo buscou incorporar a ideia de qualidade, coesão e uniformidade - e não somente eficiência - para entender melhor os precedentes resolutivos de demandas reais, ajuizadas por cidadãos brasileiros que se socorreram ao judiciário para terem uma dor sanada. Com o apoio das ferramentas tecnológicas, foi possível observar coesão de assunto de acórdãos incluídos dentro de um mesmo cluster e, a partir de uma análise manual com enfoque jurídico, foram encontrados acórdãos em que as decisões podem ser apontadas como divergências jurisprudenciais dignas de re- toques por parte da Corte Estadual. As decisões divergentes foram encontradas com maior

facilidade do que seria possível, mesmo para um especialista, caso tivesse que olhar o contingente completo de processos. O auxílio da análise construída está na possibilidade de tornar um problema relevante, mas que facilmente se perde entre os números, possível de ser encontrado e corrigido.

Referências

- Brasil (2015). *Código de Processo Civil*. Senado Federal. Lei No 13.105, de 16 de março de 2015. Edição atualizada.
- Brasil, C. N. d. J. (2023). Justiça em Números - 2023. Disponível em <https://t.ly/RVMXs>. Acessado em 05/08/2024.
- Ciurlino, V. H. (2021). BertBR : a pretrained language model for law texts.
- Didier, F. (2019). *Curso de Direito Processual Civil, Vol. 1: Introdução ao Direito Processual Civil, Parte Geral e Processo de Conhecimento*. JusPODIVM, 21 edition.
- Gomes, T. A. (2021). Avaliação de técnicas de similaridade textual na uniformização de jurisprudência. Disponível <https://repositorio.unb.br/handle/10482/40798>. Acessado em 08/06/2024.
- Lima, João Pedro e Costa, J. A. (2022). Comparing clustering techniques on brazilian legal document datasets. In *Hybrid Artificial Intelligent Systems*, pages 98–110, Cham. Springer International Publishing.
- Maaten, Laurens van der e Hinton, G. (2008). Visualizing data using t-sne. *Journal of machine learning research*, Vol. 9:p. 2579 – 2605.
- Magalhães, Dimmy; Pozo, A. e. M. S. (2023). Técnicas de aprendizado de máquinas aplicadas à classificação de decisões judiciais. *Revista de Estudos Empíricos em Direito*.
- Martins, A. D. M. (2018). Agrupamento automático de documentos jurídicos com uso de inteligência artificial. Disponível <https://repositorio.idp.edu.br/handle/123456789/2635>. Acessado em 08/06/2024.
- Mikolov, Tomas; Chen, K. C. G. e. D. J. (2013). Efficient estimation of word representations in vector space.
- Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. In *Proceedings of the International Conference on Learning Representations (ICLR)*.
- Pajankar, Ashwin e Joshi, A. (2022). *Hands-on Machine Learning with Python: Implement Neural Network Solutions with Scikit-learn and PyTorch*. Apress, Berkeley, CA.
- Polo, F. M., Mendonça, G. C. F., Parreira, K. C. J., Gianvechio, L., Cordeiro, P., Ferreira, J. B., de Lima, L. M. P., do Amaral Maia, A. C., and Vicente, R. (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. SBC.
- Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65.
- Sidorov, Grigori e Pinto, D. (2014). Soft similarity and soft cosine measure: Similarity of features in vector space model. *Computación y Sistemas*, pages 324 – 344.
- Wilton, Pablob Vigneaux e Rover, A. J. (2022). Clustering of brazilian legal judgments about failures in air transport service: an evaluation of different approaches. *Artificial Intelligence and Law*, 30:21–57. Accepted: 8 April 2021 / Published online: 17 April 2021.

Xia, C., He, T., Li, W., Qin, Z., and Zou, Z. (2019). Similarity analysis of law documents based on word2vec. In *2019 IEEE 19th International Conference on Software Quality, Reliability and Security Companion (QRS-C)*, pages 354–357.

Um *Pipeline* de Pré-Processamento de Dados Textuais em Português para Análise de Redes Sociais

Livia A. dos Santos¹, Orlando B. Coelho (*in memoriam*)¹, Renata Araujo^{1,2}, Ivan Carlos A. Oliveira¹

¹Faculdade de Computação e Informática - Universidade Presbiteriana Mackenzie
São Paulo - SP - Brasil.

²Programa de Pós-Graduação em Sistemas de Informação - EACH/USP
São Paulo - SP - Brasil

liviaalabarse.santos@mackenzista.com.br, {renata.araujo, orlando.coelho,
ivan.oliveira}@mackenzie.br

Abstract. *Preprocessing is a fundamental step in processing textual data, especially when working with text analysis, data mining or machine learning. In particular, textual data from social networks offers challenges to pre-processing, mainly due to its informal structure. This article presents a pipeline to perform 9 basic processing activities to guarantee the quality and consistency of brazilian Portuguese textual data sets extracted from social networks. Tests were conducted on datasets containing 8,000, 20,000, and 60,000 tweets, demonstrating the pipeline's performance in terms of accuracy, noise reduction, and processing time.*

Resumo. *O pré-processamento é uma etapa fundamental no processamento de dados textuais, especialmente quando se trabalha com análise de textos, mineração de dados ou aprendizagem de máquina. Em particular, os dados textuais provenientes das redes sociais oferecem desafios ao pré-processamento, principalmente devido a sua estrutura informal. Este artigo apresenta um pipeline compreendendo 9 atividades básicas de processamento para garantia da qualidade e consistência de conjuntos de dados textuais em português brasileiro extraídos de redes sociais. O pipeline foi testado em conjuntos de 8 mil, 20 mil e 60 mil tweets, demonstrando sua eficácia em termos de precisão, redução de ruído e tempo de processamento.*

1. Introdução

O pré-processamento ou preparação de dados textuais é uma das etapas mais dispendiosas e importantes no ciclo de vida de aplicações de ciência e análise de dados devido a sua natureza não-estruturada. Ela é potencializada quando se utiliza textos publicados em redes sociais. O processamento de textos de redes sociais apresenta uma variedade de desafios únicos, destacando-se a presença de gírias, erros ortográficos, emojis, abreviações e a natureza informal da linguagem. Desta forma, a etapa de pré-processamento pode conter diversas fases, tais como: remoção de URLs, exclusão de *stopwords* e símbolos especiais [Kurniawan 2020].

Muitas pesquisas envolvendo a mineração de dados textuais em redes sociais apresentam o pré-processamento nos idiomas inglês, chinês, espanhol e outras línguas [Yang e Zang 2018][Shen et. al. 2019][Kurniawan 2020][Osakwe e Cortes 2021]. No Brasil, as pesquisas na área de análise de redes sociais têm abordado esses desafios,

destacando a necessidade da limpeza e pré-processamento adequados para garantir a qualidade dos dados e o bom desempenho de análises subsequentes [Garcia et. al. 2023] [Cardozo e Freitas 2021][Nascimento et. al. 2021][Medeiros e Borges 2019][Kansaon et. al. 2018][Souza et. al. 2017]. No entanto, são poucas as pesquisas que organizam e disponibilizam para uso das comunidades científicas ou de prática, seus *pipelines* de processamento. Neste artigo, apresentamos um *pipeline* de pré-processamento de dados textuais no idioma português brasileiro para o estudo e a análise de mensagens publicadas em redes sociais e, a partir disso, permitir que aplicações que fazem uso desses dados possam extrair conhecimentos mais assertivos do seu conteúdo.

O artigo está organizado em mais três seções. A Seção 2 descreve o *pipeline* proposto, as bibliotecas e ferramentas indicadas em cada fase. A Seção 3, destaca os testes realizados com o uso do *pipeline*. A Seção 4 apresenta as conclusões, propostas de melhorias e trabalhos futuros.

2. Descrição do Pipeline

O *pipeline* desenvolvido neste trabalho¹ tem o objetivo de garantir a qualidade e consistência de conjuntos de dados textuais em português brasileiro advindos de quaisquer redes sociais, mas sua aplicabilidade vai além e pode ser empregado em diferentes bases de dados textuais. Ele combina bibliotecas/ferramentas para lidar com desafios de sua manipulação:

- **NLTK (Natural language Toolkit) v: 3.8.1** (<https://www.nltk.org/>): Utilizada para processamento de linguagem natural. Oferece uma variedade de ferramentas e recursos, incluindo tokenização e *stopwords*, que são essenciais para o pré-processamento de texto.
- **Demoji v: 1.1.0** (<https://pypi.org/project/demoji/>): Empregada para lidar com emojis presentes nos dados textuais. Fornece métodos para mapear emojis para rótulos específicos.
- **Enelvo v: 0.15** (<https://pypi.org/project/enelvo/>): Ferramenta desenvolvida para normalização de textos em português, com problemas como erros ortográficos, gírias da internet e siglas.
- **Cryptography v: 42.0.5** (<https://pypi.org/project/cryptography/>): Responsável por proteger o conteúdo do dicionário de usuários “user_dict”, garantindo que os dados sensíveis não possam ser facilmente acessados ou lidos por terceiros não autorizados.

O *pipeline* de pré-processamento é composto por etapas que transformam o texto bruto em uma representação para análise subsequente. Etapas ilustradas a seguir.

1. **Substituição de Vírgulas:** todas as vírgulas são temporariamente substituídas por um rótulo (“*chavevirg*”) para evitar conflitos durante a normalização de texto com a ferramenta Enelvo relacionadas a números com vírgulas. A substituição temporária facilita a manutenção da integridade dos dados numéricos e é revertida posteriormente. Ex. **Entrada:** [“quanto foi? 5,70?”, “que triste, queria ter ido”] **Saída:** [“quanto foi? 5chavevirg70”, “que tristechavevirg queria ter ido”]

¹ <https://github.com/ciberdem/ProjetoHEIWA-FAPESP/tree/main/CuradoriaExtracaoDados>

2. **Normalização com *Enelvo*:** Utiliza a ferramenta *Enelvo* para normalizar erros ortográficos, gírias, siglas e outros aspectos do texto. Ex. **Entrada:** [“uruguau”, “desculpa qq coisa!”, “Vc eh muitooooo legal”, “Oii, To trabahlando hj”] **Saída:** [“uruguai”, “desculpa qualquer coisa”, “você é muito legal”, “oi to trabalhando hoje”]
3. **Substituição de *Emojis*:** Substitui *emojis* encontrados no texto por rótulos específicos para uniformizar sua representação. Ex. **Entrada:** ['😊', '😋', ':)', ':('] **Saída:** ['grinningface', 'facesavoringfood', 'emojipositivo', 'emojinegativo']
4. **Substituição de Usuários:** Anonimiza usuários mencionados no texto (@usuário), substituindo-os por rótulos específicos, com um dicionário de usuários já mapeados em um arquivo criptografado chamado *'user_dict.txt'* para recuperar o rótulo correspondente. Caso contrário, cria um novo rótulo e armazena no dicionário. Ex. **Entrada:** ['oi @maria', 'gostei mt de vcs @pedro @maria', 'vamos pra praia @pedro @julia @maria?'] **Saída:** ['oi @user1', 'gostei mt de vcs @user2 @user1', 'vamos pra praia @user2 @user3 @user1?']
5. **Remoção de *URLs*:** Remove *URLs* do texto. Ex. **Entrada:** ['amei essa música! https://www.youtube.com/watch?v=dQw4w9WgXcQ'] **Saída:** ['amei essa música!']
6. **Reversão da substituição de Vírgulas:** Restaura as vírgulas substituídas no início do *pipeline*. Ex. **Entrada:** [“quanto foi? 5chavevirg70”, “que tristechavevirg queria ter ido”] **Saída:** [“quanto foi? 5,70?”, “que triste, queria ter ido”]
7. **Remoção de pontuação e Caracteres Especiais:** Remove pontuação e caracteres especiais, exceto quando são partes de *hashtags*, datas ou números com vírgula. Ex. **Entrada:** [“vai ser dia 20/05?”, “que divertido!!!! #praia”, “ quanto foi? 5,70?”] **Saída:** [“vai ser dia 20/05”, “que divertido #praia”, “quanto foi 5,70”]
8. **Remoção de *Stopwords*:** (Opcional) Remove *Stopwords*, palavras que não contribuem significativamente para o significado do texto. Ex. **Entrada:** [“vou para praia hoje”, “vou parar de fazer isso”] **Saída:** [“vou praia hoje”, “vou parar fazer”]
9. **Tokenização:** (Opcional) Divide o texto em palavras, *hashtags*, datas e números com vírgula. Cria uma lista de itens separados. Ex. **Entrada:** [“vou para praia hoje”, “vou parar de fazer isso”] **Saída:** [“vou, para, praia, hoje”, “vou, parar, de, fazer, isso”]

Embora o *pipeline* proposto tenha se mostrado eficaz em vários aspectos do pré-processamento, a ferramenta *Enelvo* apresentou algumas limitações durante a normalização de textos. Em particular, identificamos que a ferramenta não lida bem com certos tipos de gírias e abreviações frequentes em redes sociais. Isso sugere a necessidade de explorar alternativas para complementar a *Enelvo*, como o *Hunspell* (<https://hunspell.github.io/>), para atender às peculiaridades do português brasileiro usado em redes sociais.

3. Uso do *pipeline*

Para avaliar a eficácia do *pipeline* de pré-processamento de textos em redes sociais, três métricas principais foram utilizadas, sendo elas:

- **Precisão da *Enelvo*,** mede o percentual de palavras que foram corretamente normalizadas. Nesta avaliação, foi utilizado um conjunto de dados contendo 100

frases com erros ortográficos e gírias, comparando os resultados obtidos pela ferramenta com os resultados esperados. A ferramenta atingiu uma precisão de 85%, indicando boa capacidade de correção, embora haja espaço para melhorias, especialmente em casos de gírias ou abreviações com mais de um significado.

- **Tempo de Processamento**, avaliou o seu desempenho em termos de tempo de processamento. Ao lidar com conjuntos contendo: a) 8 mil *tweets*, o *pipeline* foi capaz de concluir o processo em 11 minutos; b) 20 mil *tweets*, concluiu o processo em cerca de 30 minutos; e c) 60 mil *tweets*, o tempo de processamento aumentou para aproximadamente 1 hora e 42 minutos. Todos esses experimentos foram conduzidos utilizando o *Google Colab*², em sua versão gratuita, como ambiente de desenvolvimento e execução.
- **Redução de Ruído**, avaliou a sua eficiência em remover elementos indesejados, como URLs e caracteres especiais. Para isso, foi utilizado um outro conjunto de dados com 100 frases contendo esses elementos, os seus resultados foram comparados com o conjunto de resultados esperados após a remoção de ruídos. O *pipeline* apresentou um desempenho notável, alcançando uma taxa de acerto de 97% dos caracteres especiais e 100% das URLs.

Conduzimos testes em um *dataset* composto por cerca de 20 mil postagens em português brasileiro obtidas do *Twitter (X)*. O *pipeline* foi aplicado sequencialmente a cada postagem, seguindo a ordem previamente descrita. Durante o processo, ele conseguiu lidar de forma eficaz com os desafios comuns encontrados em dados textuais em português de redes sociais. Correções ortográficas foram aplicadas, emojis foram substituídos por etiquetas específicas, menções de usuários foram anonimizadas, URLs foram removidas e a pontuação e caracteres especiais foram tratados.

4. Conclusão

Neste trabalho, foi apresentado um *pipeline* para o pré-processamento de dados textuais de redes sociais em português composto de 9 fases sequenciais, sendo as duas últimas opcionais, oferecendo flexibilidade e personalização ao processo.

Os testes fizeram uso da plataforma *Google Colab* na sua versão gratuita, sem o uso de recursos de processamento paralelo, em *datasets* com postagens do *Twitter* com quantidades de 8 mil, 20 mil e 60 mil *tweets*, com resultados satisfatórios em relação à qualidade dos dados observados, consistência e tempo de execução, apontando o potencial do *pipeline* para tarefas de análise de texto em português brasileiro. Como trabalho futuro, é possível pensar em considerar as particularidades dos textos escritos por usuários de redes sociais [Di Felippo et. al. 2021][Sanguinetti et. al. 2020]

Na construção do *pipeline*, a ferramenta Enlvo apresentou algumas falhas na normalização do texto, mostrando que uma investigação de outra tecnologia ou alteração do seu código interno pode ser adequada. A realização de mais testes envolvendo *datasets* de redes sociais com diferentes quantidades de *posts* e qualidade de conteúdo, fazendo uso de métricas (como, percentual de acertos/falhas por fase), fornecerão subsídios para avaliar com maior critério a qualidade. Testes com processamento paralelo permitirão avaliar se há melhora no seu tempo de execução.

² <https://colab.research.google.com/>

Agradecimentos

Os autores agradecem à FAPESP pelo financiamento desta pesquisa (#2021/14772-1). Renata Araujo é bolsista de produtividade em desenvolvimento tecnológico e extensão inovadora do CNPq (#305645/2022-6). Vitor dos Santos é bolsista TT1 pela FAPESP (#2023/04752-9). Livia Alabarse dos Santos é bolsista TT1 pela FAPESP (2023/04042-1).

Referências

- Cardozo, L. S. e Freitas, L. A. (2021) “Análise de Sentimentos: Avaliando o Desempenho de Pré-Processamento e de Algoritmos de Aprendizagem de Máquina sobre o Dataset TweetSentBR”, Em: *Brazilian Workshop on Social Network Analysis and Mining*. Evento Online. Sociedade Brasileira de Computação. p. 169-174.
- Di Felippo, A., Postali, C., Ceregatto, G., Gazana, L. S., Silva, E. H., Roman, N. T., Pardo, T. A. S. (2021). Descrição preliminar do corpus dantestocks: Diretrizes de segmentação para anotação segundo universal dependencies. In: *Anais do XIII Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*. SBC.
- França, T. C. e Oliveira, J. (2014) “Análise de Sentimento de Tweets Relacionados aos Protestos que ocorreram no Brasil entre Junho e Agosto de 2013”. Em: *Brazilian Workshop on Social Network Analysis and Mining*. Brasília. Sociedade Brasileira de Computação. p. 128-139.
- Garcia, L. Q., Chinellato, M. H., Caseli, H. M., Oliveira, L. H. M. (2023) “Pipeline para identificação de erros lexicais e geração de sugestões de correção”. Em: *Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana (STIL)*. Belo Horizonte/MG. Sociedade Brasileira de Computação. p. 357-361.
- Kansanon, D. P., Brandão, M. A. e Pinto, S. A. P. (2018). “Análise de Sentimentos em Tweets em Português Brasileiro” Em: *Brazilian Workshop on Social Network Analysis and Mining*. Natal. Sociedade Brasileira de Computação, 2018.
- Kurniawan, S., Gata, W., Puspitawati, D.A., Parthama, I.K.S, Setiawan, H. e Hartini, S. (2020) “Text Mining Pre-Processing Using Gata Framework and RapidMiner for Indonesian Sentiment Analysis” Em: *IOP Conference Series: Materials Science and Engineering*. IOP Publishing. p. 012057.
- Medeiros, M. C. e Borges, V. R. P.(2019) “Tweet Sentiment Analysis Regarding the Brazilian Stock Market” En: *Brazilian Workshop on Social Network Analysis and Mining*. Belém. Sociedade Brasileira de Computação. p. 71-82.
- Nascimento, R. S., Santos, G., Carvalho, F e Guedes, G. (2021) “Avaliando contribuições na substituição de termos informais em classificação de texto de redes sociais com NetSpeak-BR”. Em: *Brazilian Workshop on Social Network Analysis and Mining*. Evento Online. Sociedade Brasileira de Computação. p. 181-186.
- Osakwe, Z. T. e Cortés, Y. I. (2021) “Impact of COVID-19: a text mining analysis of Twitter data in Spanish language” Em: *Hispanic Health Care International*, v. 19, n. 4, p. 239-245.

- Sanguinetti, M., Bosco, C., Cassidy, L., Çetinoğlu, Ö., Cignarella, A. T., Lynn, T., Rehbein, I., Ruppenhofer, J., Seddah, D. e Zeldes, A. (2020). “Treebanking user-generated content: a UD based overview of guidelines, corpora and unified recommendations”. Available in: <https://arxiv.org/abs/2011.02063>.
- Shen, C., Chen, M. e Wang, C. (2019) “Analyzing the trend of O2O commerce by bilingual text mining on social media” Em: *Computers in Human Behavior*, v. 101, p. 474-483.
- Souza, B. Á., Almeida, T. G., Menezes, A. A., Figueired, C. M. S., Nakamura, F. G. e Nakamura, E. F. (2017) “Uma Abordagem para Detecção de Tópicos Relevantes em Redes Sociais Online” En: *Brazilian Workshop on Social Network Analysis and Mining*. São Paulo. Sociedade Brasileira de Computação.
- Yang, Sidi, Zhang, Haiyi. (2018) “Text mining of Twitter data using a latent Dirichlet allocation topic model and sentiment analysis” Em: *International Journal of Computer and Information Engineering*, v. 12, n. 7, p. 525-529.

Especulação Mística. Uma abordagem de Clusterização e Busca Semântica na aproximação de preço em cartas de Magic: The Gathering

Rodrigo Marques Duarte¹, André de Lima Salgado², Paula Figueira Cardoso³

¹Departamento de Automática – Universidade Federal de Lavras (UFLA)
Lavras, MG – Brasil

²Departamento de Ciência da Computação – Universidade Federal de Lavras (UFLA)
Lavras, MG – Brasil

³Faculdade de Computação – Universidade Federal do Para (UFPA)

rodrigo.duarte2@estudante.ufla.br, andre.salgado@ufla.br, pcardoso@ufpa.br

Abstract. *Magic: The Gathering (MTG) is a collectible card game that combines visual and textual elements. The release of a new collection has a direct impact on the game, leading to uncontrolled speculation on the prices of the new cards. This article presents an innovative approach to price speculation in MTG cards using clustering algorithms and semantic search. The technique allows for predicting initial prices with minimal information upon the card release and shows effectiveness in forecasting price ranges. It can also be applied to the stock market to predict the impact of news and used to prevent fraud and inflated prices by comparing existing samples.*

Resumo. *Magic: The Gathering (MTG) é um jogo de cartas colecionáveis que combina elementos visuais e textuais. O lançamento de uma nova coleção gera impactos diretos no jogo, e com isso especulações nos preços das novas cartas são tomadas sem menor controle. Este artigo apresenta uma abordagem inovadora para a especulação de preços em cartas de MTG utilizando algoritmos de clusterização e busca semântica. A técnica permite prever preços iniciais com informações mínimas no lançamento das cartas e mostra eficácia na previsão de faixas de preços, podendo ser aplicada a mercado de ações a fim de prever impacto de notícias e usada para prevenir golpes e preços exacerbados ao comparar amostras já existentes.*

1. Introdução

“*Magic: The Gathering (MTG)*” é um dos jogos de cartas colecionáveis mais antigos em existência [Guinness World Records 2024], com uma história que abrange três décadas e uma base de jogadores que ultrapassa 10 milhões [Draftsim 2024]. Diversos estudos já exploraram o MTG e seu mercado especulativo, como os realizados por [Pawlicki et al. 2014, Fink et al. 2015, Weber 2021]. Contudo, há um aspecto menos abordado: o MTG é, em sua essência, um jogo onde a linguagem desempenha um papel central, seja ela visual ou literária. Entender como essas informações são comunicadas e processadas pode oferecer insights valiosos para o Processamento de Linguagem Natural

(NLP). Com o lançamento de novas coleções, cartas têm seu preço especulado sem controle e agente controlador. Para solucionar este problema, um algoritmo para comparação da carta em lançamento com cartas já existentes foi desenvolvido. Possibilitando alertar sobre uma possível escolha alternativa já existente de menor custo assim como prevenir precificações abusivas. Neste trabalho, abordamos a problemática da classificação de preço utilizando um algoritmo de clusterização e busca semântica com o auxílio de ferramentas de NLP e modelos de linguagem. Na Seção 1, descreve-se a metodologia e recursos utilizados. Na Seção 3, apresentam-se os resultados, e por fim, na Seção 4, tecemos as considerações finais.

2. Metodologia

Para validar a proposta apresentada, foram retiradas três faixas de preço de 26.079 cartas por meio de um algoritmo de extração de conteúdo do site [Ligamagic 2024]. Essas faixas de preço foram utilizadas para construir a base de dados, que foi complementada com dados fornecidos por [Scryfall 2024]. Após a coleta, os dados passaram por um processo de limpeza para remover cartas com, custo de mana, poder e resistência dependentes de mecânicas do jogo. Na Figura 1, apresenta-se a modelagem do problema.

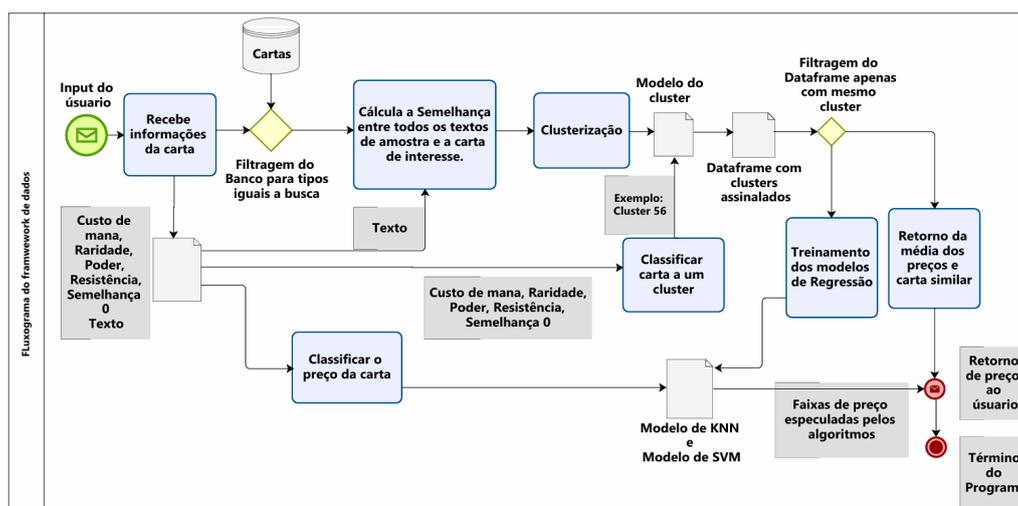


Figura 1. Fluxograma do Programa

A modelagem resultou em um algoritmo de recuperação de informação aumentada, que filtra o dataset e clusteriza cartas com atributos e textos semelhantes, permitindo a especulação de preços desde o momento de lançamento da carta, mesmo com informações mínimas [Fink et al. 2015]. A seguir, descrevemos o funcionamento do algoritmo.

2.1. Entrada de Busca, Seleção e Cálculo de Semelhanças

Inicialmente, o usuário fornece informações conhecidas sobre a carta no momento de seu lançamento, como custo de mana, raridade, poder, resistência, texto, pontos de lealdade, entre outros. Essas informações são usadas para criar um objeto que será passado pelos modelos subsequentes. O dataset é filtrado para conter apenas cartas do mesmo tipo da carta de interesse. Em seguida, calcula-se a similaridade por meio da métrica de Jaccard.

2.1.1. Métrica de Jaccard

A métrica de Jaccard é uma medida de similaridade entre dois conjuntos, calculada como o tamanho da interseção dos conjuntos dividido pelo tamanho da sua união [Manning et al. 2008]. É amplamente utilizada em problemas de comparação de textos e na análise de similaridade entre documentos. A fórmula da métrica de Jaccard é dada por:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} \quad (1)$$

O valor de $J(A, B)$ varia entre 0 e 1, onde 1 indica que os conjuntos são idênticos, e 0 indica que não há elementos em comum.

2.1.2. SentenceTransformer e Mineração de Paráfrase

A biblioteca *SentenceTransformer* é uma ferramenta para a criar representações vetoriais de sentenças que preservam relações semânticas. Um dos principais usos dessa biblioteca é na mineração de paráfrases, que envolve a identificação de sentenças que transmitem ideias semelhantes, mesmo quando formuladas de maneiras diferentes. Para realizar essa tarefa, o *SentenceTransformer* utiliza modelos pré-treinados como *BERT* e *RoBERTa* para converter sentenças em vetores de alta dimensão [Liu et al. 2019, Devlin et al. 2019]. Essas representações vetoriais são, então, comparadas entre si utilizando medidas de similaridade, como a similaridade de cosseno, para determinar o quão semanticamente próximas duas sentenças estão. O fluxo básico de trabalho envolve:

1. **Tokenização:** A sentença é dividida em tokens e convertida em uma sequência de embeddings.
2. **Transformação:** A sequência de tokens é passada através de camadas do modelo (e.g., BERT), resultando em uma representação vetorial densa da sentença.
3. **Mineração de Paráfrase:** As representações vetoriais das sentenças são comparadas usando similaridade de cosseno para identificar pares de sentenças que são paráfrases.

Em seguida, é calculada a similaridade por métrica de Jaccard entre a carta de interesse e as cartas no dataset filtrado, utilizando vetores gerados pelo modelo *SentenceTransformers*. Essa métrica de similaridade é crucial para a clusterização, que agrupa cartas com textos semelhantes conforme exemplo a seguir: [Scryfall 2024b, Scryfall 2024a]

1. “*Whenever one or more other Rabbits, Bats, Birds, and/or Mice you control enter, scry 1. Other Rabbits, Bats, Birds, and Mice you control get +1/+1.*”;
2. “*Other Squirrels you control get +1/+1.*”.

A análise dos textos revelou uma similaridade de 0,005679 entre as sentenças, além de um padrão claro nas paráfrases, como o aumento de poder e resistência de certas criaturas, onde a semântica é semelhante. Incorporar esse algoritmo de busca semântica ao *K-means* pode aprimorar o desempenho do algoritmo, pois permitirá uma clusterização mais eficiente, agrupando melhor as cartas com textos mais semelhantes.

2.1.3. Clusterização e Classificação de Pertencimento da Carta de Entrada

A clusterização visa agrupar cartas com textos semelhantes, utilizando a métrica de Jaccard como parâmetro. Para isso, foi usada a biblioteca do *sklearn* (*sklearn.clustering* e *sklearn.pipeline*) para construir a *pipeline* de clusterização. Os dados foram normalizados usando o algoritmo *MinMax* para otimizar o cálculo de distâncias no *K-means*. Foram definidos 170 clusters para prevenir a formação de agrupamentos extensos e reduzir os efeitos de *overfitting*. O coeficiente de silhueta, usado para validar a disposição dos centróides, foi de 0,6514. Com o modelo treinado, a carta de interesse é atribuída ao cluster mais adequado, filtrando o dataset para usar apenas amostras desse cluster na especulação do preço.

3. Resultados

Com base no cluster final, podemos retornar a média dos preços na amostra e determinar faixas de preço para a carta de interesse. Além disso, identificamos a carta com texto mais semelhante na amostra para verificar se já existe uma ocorrência similar do texto. As amostras são passadas por dois modelos de regressão: *Random Forest Tree (RFT)* e *KNearest Neighbor (KNN)*. Com uma divisão de 20% para teste e validação, os erros médios quadráticos e absolutos, assim como as faixas de preço previstas, são apresentados nas Tabelas 1 e 2.

RFT %	KNN %
3.672582	15.733944
78.898198	79.506967
53.006127	80.778658

Tabela 1. Erro Médio Quadrático do RFT e KNN

RFT	KNN
12.162827	10.165323
60.973797	53.930584
56.188981	55.616094

Tabela 2. Erro Médio Absoluto do SVM e KNN

Preços RFT	Preços KNN
5.41250	0.256667
13.40635	2.203333
26.90155	8.093333

Tabela 3. Faixas de preço mínimo, médio e maior preço obtida para uma carta pelos 2 algoritmos

4. Considerações Finais

Este estudo apresentou uma metodologia para a especulação de preços em cartas do MTG, combinando técnicas de NLP e algoritmos de clusterização. Os resultados demonstram que o método proposto é eficaz na previsão de faixas de preço e tem potencial para ser aplicado para prever impacto de notícias em ações e em comparação de produtos onde análises de sequências textuais são relevantes. Este trabalho foi estudo piloto para aplicação da metodologia em processos de classificação de preços de cafés populares a

partir da análise do design e conteúdo de suas embalagens. Encontramos limitações no uso da solução devido à natureza dos dados das cartas e suas altas proximidades que ocasionam centroides muito próximas. Entendemos que o uso de um modelo de vetorização com contextos treinados diretamente ao jogo pode melhorar a separação semântica das amostras.

5. Agradecimentos

Agradecemos pelo apoio e financiamento Fundação de Amparo à Pesquisa do Estado de São Paulo, FAPESP processo nº 202106968–3, além de CAPES, CNPq, UFLA, UFPA e FAPEMIG.

Referências

- Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding.
- Draftsim (2024). Mtg arena player count - how many people play? <https://draftsim.com/mtg-arena-player-count/>. Accessed: 2024-08-12.
- Fink, D., Pastel, B., and Sapra, N. (2015). Predicting the strength of magic: The gathering cards from card mechanics. *Cs 229: Machine Learning Final Project, December 2015*.
- Guinness World Records (2024). First modern trading card game. <https://www.guinnessworldrecords.com/world-records/first-modern-trading-card-game>. Accessed: 2024-08-12.
- Ligamagic (2024). Ligamagic. Acesso em: 2 ago. 2024.
- Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach.
- Manning, C. D., Raghavan, P., and Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press, Cambridge, UK.
- Pawlicki, M., Polin, J., and Zhang, J. (2014). Prediction of price increase for magic: The gathering cards. In *Proceedings of the Conference*, Stanford, CA. Stanford University.
- Scryfall (2024). Scryfall. Acesso em: 2 ago. 2024.
- Scryfall (2024a). Squirrel sovereign. Acesso em: 2 ago. 2024.
- Scryfall (2024b). Valley questcaller. Acesso em: 2 ago. 2024.
- Weber, D. (2021). Exploring markets: Magic the gathering - a trading card game. Working Paper 3/2021, IU Internationale Hochschule.

Comparação de Ferramentas para Análise de Sentimentos Aplicada no Contexto Educacional

Benjamin G. Moreira¹, Luiz C. Camargo², Ricardo J. Pfitscher¹, Tatiana R. Garcia¹

¹Universidade Federal de Santa Catarina (UFSC)
Joinville – SC – Brasil

²Centro Universitário Católica de Santa Catarina
Joinville – SC – Brasil

benjamin.moreira@ufsc.br, lzcamargo@outlook.com

ricardo.pfitscher@ufsc.br, tatiana.garcia@ufsc.br

Abstract. *This article is part of a project to mitigate dropout in programming courses in Brazilian higher education, using sentiment analysis combined with psycho-pedagogical methods and Natural Language Processing (NLP) techniques. The present research compares the effectiveness of three automated sentiment extraction tools - two based on Large Language Models (LLMs) and a lexical analyzer - using a database of 540 student responses.*

Resumo. *Este artigo faz parte de um projeto voltado para combater a evasão em disciplinas de programação no ensino superior brasileiro, utilizando análise de sentimentos combinada com métodos psicopedagógicos e técnicas de Processamento de Linguagem Natural (PLN). A presente pesquisa compara a eficácia de três ferramentas automatizadas de extração de sentimentos, duas baseadas em Large Language Models (LLMs) e um analisador léxico, usando uma base de 540 respostas de estudantes.*

1. Introdução

A análise de sentimentos é o estudo computacional das opiniões, atitudes e emoções das pessoas em relação a uma entidade. A entidade pode representar indivíduos, eventos ou tópicos. Normalmente a análise de sentimentos está associada à uma opinião sobre uma entidade, expresso em uma revisão sobre a mesma [Medhat et al. 2014]. No contexto estudantil, a análise de sentimentos tem sido vista com potencial de aplicação em diversos contextos, seja para capturar a satisfação ou o *feedback* dos estudantes sobre um determinado curso [Rani and Kumar 2017, Neumann and Linzmayer 2021], seja para compreender as dificuldades dos estudantes [Atiq and Loui 2022], ou até para mitigar potenciais casos de desistência ou evasão [Bóbbó et al. 2022].

Este trabalho é parte de um projeto de pesquisa com a finalidade de investigar a utilização do sentimento de estudantes com vistas ao uso de ferramentas automatizadas para apoio à permanência estudantil. A pesquisa utiliza a análise de sentimentos coletados de forma ativa dos estudantes, combinando métodos psicopedagógicos e técnicas automatizadas baseadas em Processamento de Linguagem Natural (PLN).

Há diferentes opções na literatura para extração dos sentimentos através de PLN. As formas mais consolidadas consistem na aplicação de analisadores léxicos, modelos

supervisionados de classificação treinados com bases públicas de textos de sentimentos (normalmente com bases de avaliações sobre produtos e filmes, como é o caso do IMDb) [Wankhade et al. 2022]. Por outro lado, o crescente interesse nas ferramentas de inteligência artificial generativas, que utilizam de Large Language Models (LLMs) para construção de conhecimento [Chang et al. 2024], tem instigado pesquisadores de diversas áreas a aplicar essas ferramentas na extração de sentimentos [Mughal et al. 2024].

2. Trabalhos relacionados

A análise de sentimentos é realizada em diversos domínios de aplicação. Em [Lazarini et al. 2023, Seno et al. 2023], o objetivo é entender melhor o sentimento público em debates políticos. No contexto educacional, em [Atiq and Loui 2022], é apresentado um estudo sobre as emoções de estudantes durante a realização de atividades de programação, com observações qualitativas. Em [Bóbó et al. 2022], a análise de sentimentos é utilizada para prever o risco de evasão com dados coletados de textos presentes em um ambiente virtual de aprendizagem. Ainda no domínio educacional, um mapeamento sistemático é apresentado em [Coto et al. 2022].

Outro trabalho próximo a esse artigo é apresentado em [Mughal et al. 2024], em que uma extensiva análise é realizada para comparar ferramentas de aprendizado profundo e LLMs quanto ao desempenho na análise de sentimentos baseada em aspectos (ABSA, do inglês aspect-based sentiment analysis). Os resultados da avaliação de bases de dados públicas de domínios específicos (opiniões sobre hotéis, restaurantes e livros) mostram que o modelo PaLM apresentou os melhores resultados na maioria dos casos, sendo inferior ao GPT3.5 em uma base de dados projetada para conter ao menos dois sentimentos em uma mesma frase.

3. Metodologia

Os dados utilizados neste trabalho foram obtidos a partir da coleta ativa de sentimentos de alunos. Detalhes do processo são apresentados em [Pfitscher et al. 2023]. A coleta foi realizada por meio de um questionário respondido de forma anônima. Duas perguntas foram realizadas: (1) Como você está se sentindo em relação à disciplina?; e (2) Como você está se sentindo em relação à universidade?

Cada resposta recebeu uma classificação psicopedagógica, realizada de modo manual por duas psicopedagogas, que empiricamente buscaram distinguir termos de sentimentos dos demais em cada frase das respostas dos estudantes. Para esta classificação, primeiramente as avaliadoras definiram as categorias e estabeleceram as listas de itens sobre os aspectos positivos e negativos na experiência com a disciplina/universidade utilizando a análise de conteúdo [Bardin 1977], que é uma análise qualitativa do discurso manifesto da comunicação. Foi estabelecido um conjunto de categorias de classificação para orientar esse processo:

- *Positivo*: respostas exclusivamente positivas;
- *Negativo*: respostas exclusivamente negativas;
- *Ambos*: respostas com menções positivas e negativas;

A classificação como *Ambos* mostrou-se necessária uma vez que diversas respostas são formadas por diversas frases, podendo indicar tanto aspectos positivos quanto negativos dos itens avaliados.

Para uma classificação automática dos sentimentos, foram comparados três modelos, dois deles baseados em LLMs (ambas as LLMs são opções open source) e outro com análise baseada em Léxico. Segue descrição dos modelos:

- Llama¹ (versão Meta Llama 3 Instruct I 8B): O Llama 3 é um LLM desenvolvido pela empresa Meta. A versão utilizada é o modelo com 8 bilhões de parâmetros (o menor disponível).
- Gemma² (versão Gemma 2 9B IT): O Gemma 2 é o LLM de código aberto do Google. A versão utilizada é o modelo com 9 bilhões de parâmetros.
- LeIA³ (Léxico para Inferência Adaptada): é uma adaptação para o português do léxico e ferramenta para análise de sentimentos VADER (Valence Aware Dictionary and sEntiment Reasoner).

Para as LLMs, foi utilizado como *prompt* do sistema a seguinte entrada: “Análise o sentimento da frase e responda apenas com POSITIVO, NEGATIVO ou AMBOS”. Também foi fornecida a temperatura de respostas com o valor zero⁴.

Para utilização do LeIA, foi definido que o parâmetro *compound* superior à 0,05 foi considerado *Positivo*, um *compound* inferior à -0,05 foi considerado *Negativo*, e o intervalo utilizado para definir o sentimento como *Ambos*. Essa faixa da classificação como *Ambos* é considerada para uma pontuação neutra, na qual existe um equilíbrio entre termos positivos e negativos, mas foi considerada nesse trabalho como uma possibilidade para qualificar a existência das duas polaridades no mesmo texto.

4. Resultados

Nesta seção, são apresentados os resultados dos modelos aplicados em comparação com as classificações psicopedagógicas realizadas.

Os modelos obtiveram acurácia aproximada de 75%, 77% e 62% para o Llama, Gemma e LeIA, respectivamente. Embora o Gemma tenha obtido melhor acurácia, uma análise a partir das matrizes de confusão (conforme a Figura 1), mostra que o Gemma diminuiu o acerto para sentimentos positivos e negativos. Por outro lado, o Llama teve melhor desempenho em classificar os sentimentos positivos e negativos, mas praticamente não conseguiu classificar os sentimentos definidos como *Ambos*, classificando esses, em sua maioria, como negativos. Já o modelo LeIA teve uma taxa de acertos na classificação de sentimentos positivos e negativos similar ao Gemma, mas a classificação para a categoria *ambos*, mostrou-se bastante ruim.

Aprimorando a análise e, uma vez que as classes estão desbalanceadas, é considerado o resultado da métrica F1-score, que apresenta uma medida mais robusta do desempenho do modelo, levando em consideração tanto a precisão quanto a revocação nas classificações. Os valores percentuais apresentados nas discussões sobre o F1-score são apresentados de forma aproximada.

¹Site do projeto do Llama: <https://llama.meta.com>

²Site do projeto do Gemma: <https://ai.google.dev/gemma>

³Site do projeto do LeIA: <https://github.com/rafjaa/LeIA>

⁴Utiliza-se uma temperatura baixa para que a LLM forneça respostas mais precisas, o que é útil quando se deseja consistência e precisão.

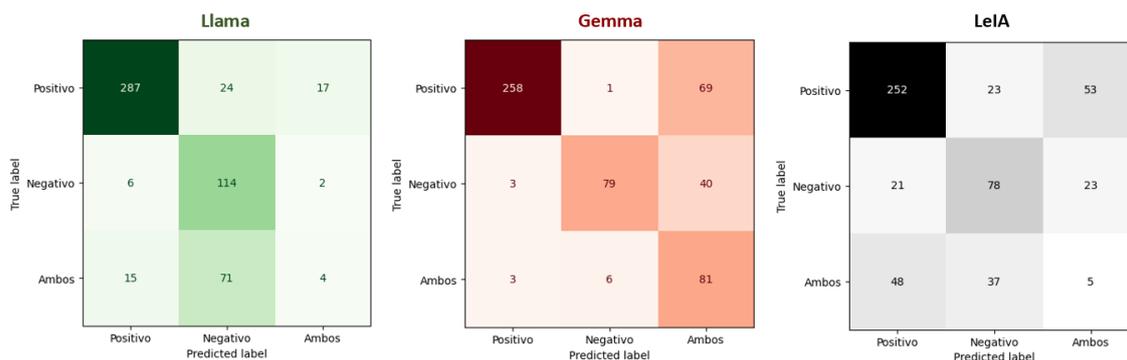


Figura 1. Matrizes de confusão da classificação dos sentimentos

Os três modelos obtiveram F1-score considerado bom na classificação de sentimentos positivos, sendo que a Llama obteve acerto de 90%, enquanto o Gemma obteve 87% e o LeIA obteve 77%. A qualidade da classificação realizada pelos modelos diminui quanto aos sentimentos negativos, sendo que o Llama obteve 69%, Gemma alcançou 76% e LeIA obteve 60%. Observa-se que, embora o Llama tenha obtido uma melhor classificação dos sentimentos positivos, para o sentimento negativo o F1-score é reduzido significativamente (21% do Llama, contra 11% do Gemma).

Para classificação como *Ambos*, o Llama obteve 71%, Gemma 58% e LeIA apenas 6%. O Llama teve desempenho similar na classificação de *Ambos* quanto nos sentimentos negativos, enquanto o Gemma teve uma redução significativa. Quanto ao LeIA, sua classificação para a categoria *Ambos* é insignificante. Esses valores representam aspectos percebíveis a partir das matrizes de confusão, onde existe um “espalhamento” das classificações de sentimentos negativos e ambos.

5. Considerações finais

Neste artigo, dados coletados ativamente de estudantes de programação foram utilizados na extração e análise de sentimento estudantil. Sob a perspectiva da métrica F1-score, os resultados obtidos com as ferramentas são considerados bons com a classificação de sentimentos “positivos”. Para os sentimentos “negativos”, a qualidade da classificação não é mantida e há um declínio de desempenho das três ferramentas. O desempenho segue em declínio para as ferramentas Llama e LeIA na classificação “Ambos”, mas não para a ferramenta Gemma, que manteve o nível desempenho para esse tipo de classificação próximo do obtido com a classificação “negativos”.

Uma vez que a classificação como *Ambos* possui menor acerto, uma alternativa é classificar as respostas por frases, indicando quantos sentimentos positivos e negativos estão presentes na resposta. Como a classificação de sentimentos como positivos e negativos apresentou boa taxa de acertos, o resultado da classificação pode ser mais confiável.

Para a aplicação pretendida, que envolve a permanência estudantil, é considerado como mais importante a identificação dos sentimentos negativos. Dessa forma, a escolha do Llama é a mais indicada, uma vez que essa identifica melhor os sentimentos negativos e positivos, bem como as classificações incorretas dos *Ambos* serem, em sua maioria, classificadas como negativos (entende-se que os aspectos negativos podem ter potencial em impactar na relação do estudante com a disciplina).

Referências

- Atiq, Z. and Loui, M. C. (2022). A qualitative study of emotions experienced by first-year engineering students during programming tasks. *ACM Transactions on Computing Education (TOCE)*, 22(3):1–26.
- Bardin, L. (1977). *Análise de conteúdo*. Lisboa: edições 70.
- Bóbbó, M. L., Campos, F., Stroele, V., David, J. M. N., Braga, R., and Torrent, T. T. (2022). Using sentiment analysis to identify student emotional state to avoid dropout in e-learning. *International Journal of Distance Education Technologies (IJDET)*, 20(1):1–24.
- Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., Chen, H., Yi, X., Wang, C., Wang, Y., et al. (2024). A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 15(3):1–45.
- Coto, M., Mora, S., Grass, B., and Murillo-Morera, J. (2022). Emotions and programming learning: systematic mapping. *Computer Science Education*, 32(1):30–65.
- Lazarini, L., Anno, F. S. I., Seno, E. R. M., and Caseli, H. M. (2023). Abordagens baseadas em léxicos para a classificação de sentimentos orientada aos alvos de opinião em comentários do domínio político. In *Anais do XIV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 375–380. SBC.
- Medhat, W., Hassan, A., and Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams engineering journal*, 5(4):1093–1113.
- Mughal, N., Mujtaba, G., Shaikh, S., Kumar, A., and Daudpota, S. M. (2024). Comparative analysis of deep natural networks and large language models for aspect-based sentiment analysis. *IEEE Access*, 12:60943–60959.
- Neumann, M. and Linzmayer, R. (2021). Capturing student feedback and emotions in large computing courses: A sentiment analysis approach. In *Proceedings of the 52nd ACM Technical Symposium on Computer Science Education*, pages 541–547.
- Pfitscher, R., Camargo, L., Moreira, B., Wang, C., Zedral, R., and Garcia, T. (2023). Análise de sentimentos em turmas de programação com vistas ao apoio à permanência estudantil. In *Anais do XXXIV Simpósio Brasileiro de Informática na Educação*, pages 1329–1340, Porto Alegre, RS, Brasil. SBC.
- Rani, S. and Kumar, P. (2017). A sentiment analysis system to improve teaching and learning. *Computer*, 50(5):36–43.
- Seno, E. R. M., Anno, F. S. I., Lazarini, L., and Caseli, H. M. (2023). Classificação de polaridade orientada aos alvos de opinião em comentários sobre debate político em português. In *Anais do XIV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 84–93. SBC.
- Wankhade, M., Rao, A. C. S., and Kulkarni, C. (2022). A survey on sentiment analysis methods, applications, and challenges. *Artificial Intelligence Review*, 55(7):5731–5780.

Modelo de Linguagem Quantizados na Área da Saúde: Um Enfoque em Perguntas e Respostas com Base na Técnica DPO

Mario Pinto Freitas Filho, João Dallyson Sousa de Almeida, Anselmo C. Paiva

¹Universidade Federal do Maranhão (UFMA)

²NCA - Nucleo de Computação Aplicada

Resumo. *A agilidade em diagnosticar os pacientes é um fator vital para o tratamento hábil de diversas enfermidades e muitas vezes é o parâmetro decisivo na recuperação dos pacientes. Ao levar em consideração que o tempo médio consumido por profissionais médicos em atividades de pesquisa, muitas vezes é de 4 horas de duração, e este se reduzido, de forma a não comprometer a qualidade dos resultados obtidos será valioso para o diagnóstico e tratamento principalmente em casos de maior urgência. A presente produção busca explorar a utilização de Modelos dos Large Language Models (LLMs) baseados na arquitetura Transformer para otimizar o tempo e a eficiência nas atividades de pesquisa de profissionais de saúde. Para tanto, objetiva-se compreender o que são as LLM através do Transformer e suas funcionalidades além de apresentar o dataset-Medtext utilizado para treinar o modelo. Portanto, esta produção trata-se de uma pesquisa experimental na qual será aplicado o conhecimento teórico sobre LLMs e Transformers para resolver a problemática e otimização do tempo de pesquisa.*

Abstract. *The agility in diagnosing patients is a vital factor for the skillful treatment of various diseases and is often the decisive parameter in patient recovery. Considering that the average time spent by medical professionals on research activities is often 4 hours, reducing this time without compromising the quality of the results would be valuable for diagnosis and treatment, especially in more urgent cases. This work aims to explore the use of Large Language Models (LLMs) based on the Transformer architecture to optimize the time and efficiency of healthcare professionals' research activities. To this end, the objective is to understand what LLMs are through the Transformer and their functionalities, in addition to presenting the Medtext dataset used to train the model. Therefore, this work is an experimental study in which theoretical knowledge about LLMs and Transformers will be applied to address the problem and optimize research time.*

1. Introdução

O avanço acelerado das técnicas de Processamento de Linguagem Natural (PLN) nos últimos anos, resultou no surgimento de modelos de chat altamente sofisticados, como o GPT-4, LLAMA 2, e Falcon. Esses modelos exibem uma notável capacidade de compreender e gerar respostas semelhantes às humanas em diversos domínios, tornando-os cada vez mais populares em aplicações como suporte ao cliente, assistentes virtuais, moderação de mídia social, entretenimento e pesquisa.

No entanto, apesar do seu potencial, esses modelos são frequentemente acessíveis apenas por meio de APIs restritas, o que limita novas pesquisas e avanços na área de PLN. Além disso, eles são extremamente pesados e exigem grande poder computacional para treinamento e inferência.

Com isso em mente, o presente artigo busca apresentar abordagens para reduzir a demanda computacional desses modelos extensos, conhecidos como LLMs [Chen et al. 2023]. Serão exploradas técnicas como GPTQ [Frantar et al. 2023] para quantização de modelos e DPO (*Direct Preference Optimization*) [Rafailov et al. 2023] para otimização de pesos, as quais serão discutidas nas seções subsequentes.

Um estudo recente com profissionais de saúde revelou que eles gastam, em média, 60 minutos para formular uma estratégia de busca e durante suas pesquisas e dedicam cerca de 3 minutos para avaliar a relevância de cada documento, totalizando aproximadamente 4 horas de pesquisa [Russell-Rose T 2017]. Uma solução viável para enfrentar esses desafios é a implementação de sistemas de perguntas e respostas dedicados. Esses modelos são capazes de compreender perguntas em linguagem natural e fornecer respostas baseadas em uma base de dados validada por especialistas.

A presente produção é composta pela fundamentação teórica em que são relatadas as técnicas utilizadas para a criação do modelo. A seguir são descritos o dataset utilizados, seu pre-processamento e experimentos realizados e por fim as considerações finais em que é sintetizado o trajeto percorrido até os resultados.

2. Fundamentação teórica

2.1. Arquitetura e Funcionamento de LLMs

Antes da arquitetura *Transformer*, modelos como LSTM (*Long Short-Term Memory*) e GRU (*Gated Recurrent Unit*) mitigavam, mas não resolviam completamente, a limitação dos RNNs (*Recurrent neural networks*) em lidar com dependências de longo alcance em sequências. O *Transformer*, apresentado em [Vaswani et al. 2023], introduziu uma abordagem inovadora com o mecanismo de atenção e o esquema codificador-decodificador, permitindo que o codificador transforme a sequência de entrada em uma representação vetorializada compreensível pela máquina.

Uma arquitetura *Transformer* em PLN, se configura por uma sequência de entradas passadas por uma camada de incorporação e codificação posicional antes de ser processada pelo codificador, que captura a semelhança entre as palavras e suas posições. O decodificador, então, usa esses vetores para produzir a saída de forma auto-regressiva, onde cada *tokens* de saída torna-se a entrada para o próximo passo.

O termo “auto-regressivo” refere-se ao processo de gerar saídas sequenciais, permitindo ao modelo criar frases de saída de comprimentos variáveis, adaptando-se a diferentes contextos e requisitos.

Llama-2 [Touvron et al. 2023] é uma coleção de quatro modelos baseados na arquitetura *Transformer*, variando em parâmetros: 7B, 13B, 34B e 70B. Todos compartilham a mesma função de ativação e método de normalização. Neste trabalho, foi utilizado o modelo LLAMA-2 7B.

O mesmo se diferencia por um novo método de ajuste fino chamado Ghost Atten-

tion (GAtt), projetado para que o modelo mantenha consistentemente o papel atribuído pelo usuário, como por exemplo "Cardiologista". O GAtt adiciona sinteticamente a instrução "agir como" a todas as mensagens do usuário, permitindo que o modelo mantenha o contexto sem precisar da concatenação explícita durante o ajuste fino. Essa técnica melhora o controle do diálogo em múltiplos turnos, permitindo melhor adaptação às instruções dos usuários.

2.2. Quantização do modelo (GPTQ)

devido às características do modelo LLAMA 2 7B, com 7 bilhões de parâmetros, sua execução em GPUs convencionais ou no Google Colab não foi viável por causa da alta demanda de recursos. Para mitigar essa limitação, foram exploradas técnicas de quantização para reduzir o consumo de VRAM (*Video Random Access Memory*) sem comprometer o desempenho do modelo. Dois métodos principais de quantização são mencionados: a Quantização Pós-Treinamento (PTQ), que quantiza um modelo já pré-treinado, e o Treinamento com Reconhecimento de Quantização (QAT), que realiza a quantização antes ou durante o treinamento. O GPTQ (*Generative Pretrained Transformers Quantization*), uma técnica PTQ (*Post-training quantization*), é ideal para modelos muito grandes, onde o treinamento completo seria muito custoso.

O GPTQ utiliza um esquema misto de quantização int4/fp16 (*integer* de 4 bits e fp16 *float* de 16 bits), onde os pesos são quantizados como int4 e as ativações permanecem em float16. Durante a inferência, os pesos são desquantizados instantaneamente e o cálculo é feito em float16. Esse esquema oferece dois benefícios principais: economia de memória de até 4 vezes devido à quantização int4, e potencial aceleração da inferência e do treinamento, graças à menor largura de bits utilizada para os pesos, o que reduz o tempo de comunicação de dados [Frantar et al. 2023].

3. Materiais e Métodos

3.1. Dataset

MedText [Melamud and Shivade 2019] é um conjunto de dados para diagnósticos e tratamentos médicos, contendo 1.412 perguntas e respostas baseadas em casos de pacientes, vale ressaltar que este encontra-se na língua inglesa e aborda as 100 doenças e 30 lesões mais comuns nos hospitais. Cada condição possui perguntas e respostas variando entre leve, complicada e grave.

Ele foi desenvolvido a partir do MIMIC III [Johnson et al. 2018], um banco de dados que contém informações de prontuários médicos, incluindo diagnósticos, causas e tratamentos. A partir desses dados, foi criada uma estrutura artificial de perguntas e respostas (*question answering*) para o MedText, conforme descrito em [Melamud and Shivade 2019].

3.2. Pre-processamento do dataset e DPO

O pré-processamento do dataset é muito importante para o funcionamento do DPO (*Direct Preference Optimization*), que é uma alternativa ao aprendizado por reforço com feedback humano (RLHF). O DPO melhora o alinhamento da linguagem com as preferências humanas sem a necessidade de um modelo de recompensa. Ele utiliza dados compostos por triplas (*prompt*, resposta escolhida, resposta rejeitada), também usados no RLHF. Durante

	perplexidade
[Melamud and Shivade 2019]	12.5
Este artigo	1.98

Tabela 1. Metricas

o ajuste fino, o modelo de linguagem é duplicado, criando um modelo treinável (*policy model*) e outro congelado (*reference model*), ambos responsáveis por avaliar as respostas com base nas probabilidades dos *tokens*.

O DPO possui suas próprias métricas de avaliação, como a diferença de probabilidades logarítmicas entre as respostas escolhidas e rejeitadas, e a frequência de casos em que as respostas escolhidas superam as rejeitadas. Devido à estrutura de *prompt*, resposta escolhida e rejeitada, foi necessário adaptar o MedText, utilizando a tecla (espaço) a fim de preencher campo rejeitado sem adicionar quaisquer informação que possa interferir no processo. Zephyr [Tunstall et al. 2023], um *chatbot* que compete com grandes LLMs como mostrado em [Li et al. 2023] e [Zheng et al. 2023], adotou uma abordagem semelhante, mas a presença da teclada de (espaço) pode ser um risco visto que o modelo pode gerar respostas equivocadas.

3.3. Experimentos

A métrica usada na avaliação do modelo foi a perplexidade com isso em mente realizou-se a validação do modelo desenvolvido, comparando-o com o artigo original de onde o dataset foi extraído [Melamud and Shivade 2019]. Utilizando a técnica DPO, o modelo foi treinado por 10 épocas com uma taxa de aprendizado de $5e-07$ e batch size de 2, permitindo uma análise detalhada das amostras. O otimizador AdamW foi escolhido para otimizar a eficiência na convergência. Os resultados visto na tabela 1 mostram que o modelo proposto teve uma perplexidade de **1.98**, Consideravelmente superior aos resultados do artigo do qual o conjunto de dados foi extraído [Melamud and Shivade 2019] **12.5**, indicando maior precisão na previsão de palavras e na compreensão das estruturas linguísticas.

4. Conclusão e Agradecimentos

O texto aborda estratégias para reduzir o custo computacional de grandes modelos de linguagem (LLMs), que frequentemente precisam ser executados em APIs externas. Uma das estratégias mencionadas é o GPTQ, que quantiza o modelo, reduzindo o uso de memória (VRAM), o tempo de treinamento e de inferência. O DPO, por sua vez, ao usar decodificadores causais, permite calcular recompensas em um único passo, otimizando o processo de treinamento.

Os autores agradecem Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Finance Code 001, Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), e a Fundação de Amparo à Pesquisa Desenvolvimento Científico e Tecnológico do Maranhão (FAPEMA), Empresa Brasileira de Serviços Hospitalares (Ebserrh) Brasil (Grant number 409593/2021-4) pelo financiamento.

Referências

- Chen, Z., Mao, H., Li, H., Jin, W., Wen, H., Wei, X., Wang, S., Yin, D., Fan, W., Liu, H., and Tang, J. (2023). Exploring the potential of large language models (llms) in learning on graphs.
- Frantar, E., Ashkboos, S., Hoefler, T., and Alistarh, D. (2023). Gptq: Accurate post-training quantization for generative pre-trained transformers.
- Johnson, A. E. W., Stone, D. J., Celi, L. A., and Pollard, T. J. (2018). The mimic code repository: enabling reproducibility in critical care research. *Journal of the American Medical Informatics Association*, 25(1):32–39.
- Li, X., Zhang, T., Dubois, Y., Taori, R., Gulrajani, I., Guestrin, C., Liang, P., and Hashimoto, T. B. (2023). Length-controlled alpacaEval: A simple way to debias automatic evaluators. https://github.com/tatsu-lab/alpaca_eval.
- Melamud, O. and Shivade, C. (2019). Towards automatic generation of shareable synthetic clinical notes using neural language models. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop at NAACL*, pages 35–45.
- Rafailov, R., Sharma, A., Mitchell, E., Ermon, S., Manning, C. D., and Finn, C. (2023). Direct preference optimization: Your language model is secretly a reward model.
- Russell-Rose T, C. J. (2017). Expert search strategies: The information retrieval practices of healthcare information professionals. *JMIR Med Inform* 2017;5(4):e33.
- Touvron, H., Martin, L., and Al, E. (2023). Llama 2: Open foundation and fine-tuned chat models.
- Tunstall, L., Beeching, E., Lambert, N., Rajani, N., Rasul, K., Belkada, Y., Huang, S., von Werra, L., Fourier, C., Habib, N., Sarrazin, N., Sanseviero, O., Rush, A. M., and Wolf, T. (2023). Zephyr: Direct distillation of lm alignment.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2023). Attention is all you need.
- Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E. P., Zhang, H., Gonzalez, J. E., and Stoica, I. (2023). Judging llm-as-a-judge with mt-bench and chatbot arena.

LLM-SEMREL: Towards a Better Coreference Resolution for Portuguese

Evandro Fonseca¹, Joaquim Neto¹

¹Blip

evandro.fonseca, joaquim.neto {@blip.ai}

Abstract. *This paper aims to describe LLM-SEMREL, a new Portuguese semantic database built automatically using currently available large language models. The motivation for this project stems from the lack of rich semantic resources for the Coreference Resolution task in Portuguese. As a result, we provide a new resource that can be used to improve current models and build new ones. LLM-SEMREL is composed of 1,229,399 semantic relations, distributed among 261,731 words and their descriptions.*

1. Introduction

The development of Large Language Models (LLMs) has made it possible to create large volumes of synthetic data automatically, which is particularly useful in contexts where annotated data is scarce. These models, trained on vast text corpora, can be fine-tuned to generate examples that follow specific semantic patterns, replicating the linguistic complexity required for tasks such as coreference resolution. In the context of Brazilian Portuguese, this approach has the potential to overcome the limitation of resources, providing a diverse and enriched dataset that is crucial for training machine learning models. This capability is particularly significant when we consider the complexity and challenges of coreference resolution. According to [Brown et al. 2020], LLMs like GPT-3 can generate text that mimics the intricacies of natural language, which can be leveraged to create high-quality synthetic data for various NLP tasks.

Coreference resolution is a central task in Natural Language Processing (NLP), involving the identification of all expressions in a text that refer to the same entity[Fonseca 2018]. This task is challenging due to the inherent complexity of natural language, where different forms and expressions can refer to the same concept or object. In this context, semantic databases, which categorize and interrelate meanings of words and expressions, can provide the necessary support for machine learning systems to make precise inferences about which expressions are coreferent [Khosla and Rose 2020].

In recent years, the development of LLMs has revolutionized the field of NLP[OpenAI 2023, Dubey et al. 2024, Reid et al. 2024]. Leveraging the capabilities of these advanced models, we present a comprehensive semantic

database specifically tailored for the Portuguese language. This database, created using state-of-the-art LLMs, aims to enhance various NLP applications by providing rich semantic information and nuanced understanding of Portuguese text. This initiative is particularly critical when considering the challenges associated with creating resources in Brazilian Portuguese considering the scarcity of resources for the task in Portuguese.

However, the availability of resources such as semantic databases and annotated corpora in Brazilian Portuguese is limited compared to other languages like English[Hedderich et al. 2021]. While some resources are available, they are often insufficient to cover the linguistic diversity present in Brazilian Portuguese, which hinders the development of robust coreference resolution systems.

The creation of an automated, Portuguese-specific semantic database is, therefore, of utmost importance for advancing the task of Portuguese coreference resolution. This resource would not only facilitate the training of more accurate models but could also be used in a wide range of other NLP tasks, such as machine translation and sentiment analysis. With a robust semantic database, AI systems can achieve a deeper and more contextualized understanding of the Portuguese language, enabling them to solve complex tasks more effectively. Additionally, the availability of such resources would contribute to the democratization of technology, allowing more Brazilian researchers and developers to create innovative solutions for the local market, reinforcing the relevance of initiatives in this direction.

2. Related Work

In the study conducted by [Fonseca et al. 2016], the authors explore the impact of integrating semantic knowledge into coreference resolution systems for the Portuguese language. Their work specifically evaluates the contribution of semantic features derived from the Onto.PT lexical resource, which includes relations such as synonymy, hypernymy, and hyponymy. By incorporating these semantic features into a machine learning model, they were able to enhance the model’s ability to correctly identify coreferent entities, particularly in cases where traditional lexical and syntactic features might fall short. The results showed improvements in precision, recall, and F-measure, highlighting the importance of utilizing comprehensive semantic databases to enrich coreference resolution tasks, especially in resource-limited languages like Portuguese.

[Jiang and Cohn 2021] introduced a coreference resolution model that incorporates syntactic and semantic information through a Heterogeneous Graph Attention Network (HGAT). This model constructs a heterogeneous graph that integrates syntactic structures, such as dependency trees, with semantic structures derived from Semantic Role Labeling (SRL). By enhancing word representations with this combined syntactic and semantic context,

the model significantly improves coreference resolution accuracy compared to models relying solely on pre-trained embeddings like SpanBERT. This work underscores the importance of well-structured semantic databases and syntactic inputs in enhancing the performance of coreference resolution systems.

A study conducted by [Lima et al. 2018] analyzed the use of different semantic bases, specifically Onto.PT[Gonçalo Oliveira 2012] and ConceptNet[Speer and Havasi 2012], in the task of coreference resolution for Portuguese texts. The authors demonstrated that integrating semantic information from these bases can improve the performance of coreference tools like CORP, particularly when dealing with relations of hyponymy and synonymy. The research concluded that while both semantic bases contribute to the system's performance, the use of ConceptNet resulted in slightly higher precision, highlighting the importance of semantic bases in enhancing coreference resolution, especially in languages with limited resources.

3. LLM-SEMREL

LLM-SEMREL¹, as its name suggests, was entirely annotated using a large language model. We used GPT-4o [OpenAI 2023] to annotate 261,731 words, considering eight semantic relations. The LLM-SEMREL words were taken from the br.ispell dictionary²[Ueda 2005].

Regarding our method for developing the semantic database, we used the prompt-tuning technique, specifically zero-shot learning, to develop the resource. Through the Azure API and using the GPT-4o model, we created a Java code that made a request to the API for each word in the br.ispell dictionary and executed the prompt. The response from the LLM request is a JSON in the following format: first, we have the field *word*, which contains the word from the dictionary; then there is the *description* field, which provides a general context of the meaning of the word in the *word* field; finally, we have the *relations* field, which is a list of our semantic relations. Below, we show an example of the JSON for the word "cachorro"

```

1 {
2   "word" : "cachorro",
3   "description" : "Mamífero doméstico, conhecido por sua
4     lealdade e companheirismo com os seres humanos.",
5   "relations" : [ {
6     "synonym_of" : [ "cão", "canino" ]
7   }, {
8     "hypernym_of" : [ "animal de estimação", "mamífero" ]
9   }, {
10    "hyponym_of" : [ "pastor-alemão", "poodle", "labrador" ]
11  } ]

```

¹the resource is available at: <https://github.com/evandrofonsecatake/llm-semrel>

²<https://www.ime.usp.br/pf/dicios/>

```

11     "meronym_of" : [ "pata", "cauda", "focinho" ]
12   }, {
13     "holonym_of" : [ "matilha" ]
14   }, {
15     "paronym_of" : [ "cachorrinho", "cachorrão" ]
16   }, {
17     "troponym_of" : [ ]
18   }, {
19     "antonym_of" : [ "gato" ]
20   } ]
21 }

```

Below, we present our prompt, which is applied to each word in the dictionary to generate semantic relations. We begin by establishing that the agent is responsible for generating semantic annotations when they exist for a given word. Following this, we provide detailed instructions on how the output should be formatted.

Você é um assistente que realiza anotação linguística de relações semânticas. Você precisa anotar as relações quando existentes para cada palavra recebida, seguindo o json exemplo:

```

"word":
"description":
"relations": [
"synonym_of": []
"hypernym_of": []
"hyponym_of": []
"meronym_of": []
"holonym_of": []
"paronym_of": []
"troponym_of": []
"antonym_of": []
]

```

Figure 1. Prompt used to collect the semantic relations of a word.

In Table 1, we show the relations, their definitions, and examples of each. The generated database comprises a total of 1,229,399 semantic relations.

In Table 2, we compare the number of relations between LLM-SEMREL and Onto.PT. We can see that Onto.PT has more types of relations. However, our resource was built with a focus on semantic relations to solve coreferences. In Table 3, we show the number of tokens consumed by the LLM model

Table 1.

Relation	Definition	Example	Number of Relations
synonym_of	Synonymy refers to the relationship between words that have similar or identical meanings.	feliz and alegre	447,773
antonym_of	Antonymy refers to the relationship between words that have opposite meanings.	quente and frio	335,558
hyponym_of	Hyponymy refers to the relationship where a word has a more specific meaning than a general or superordinate term.	pardal is a hyponym of pássaro	68,360
hypernym_of	Hypernymy refers to the relationship where a word has a broader meaning that encompasses more specific words.	veículo is a hypernym of carro	60,310
meronym_of	Meronymy denotes a part-whole relationship where a word represents a part of something larger.	roda is a meronym of carro	15,554
holonym_of	Holonymy is the relationship where a word represents the whole to which parts belong.	árvore is a holonym of galho	10,461
paronym_of	Paronymy refers to words that are similar in form or derivation but have different meanings.	cavalheiro and cavaleiro	274,994
troponym_of	Troponymy is the relationship where a verb denotes a specific manner of doing something that another verb denotes.	sussurrar is a troponym of falar	16,389

to generate the whole database. It is possible to see that 71,129,390 tokens were expended.

Table 2.

Relation	LLM-SEMREL	Onto.PT
Synonym_of	447,773	168,858
Antonym_of	335,558	92,598
Hyponym_of	68,360	91,466 (combined)
Hypernym_of	60,310	91,466 (combined)
Meronym_of	15,554	9,436
Holonym_of	10,461	7,431
Paronym_of	274,994	-
Troponym_of	16,389	-
Contained_In	-	644
Material_of	-	873
Cause_of	-	12,369
Producer_of	-	2,303
Purpose_of	-	16,271
Has_Quality	-	2,256
Has_State	-	561
Property_of	-	38,048
Place_of	-	1,393
Manner_of	-	3,966
Manner_Without	-	265
Total	1,229,399	448,738

4. Error Analysis

Regarding error analysis, we have examined several instances to understand the primary errors in our presented semantic relations. According to our anal-

Table 3.

Tokens			Words	Semantic Relations
Prompt	Completion	Total		
39,360,451	31,768,939	71,129,390	261,731	1,229,399

ysis, the main errors found in the database are due to the directionality between the semantic relations of hypernym, hyponym, meronym, and holonym. It is known that hyponyms are more specific than hypernyms (subclass relation), just as meronyms represent parts of something and holonyms represent the whole. To illustrate that we present two instances: “cachorro”(dog) and “Terra”(Earth):

Table 4. pathways of semantic relations

Word	Relations				
	synonym_of	hypernym_of	hyponym_of	meronym_of	holonym_of
Cachorro	cão, canino	animal de estimação, mamifero	pastor-alemão, poodle, labrador	pata, cauda, focinho	matilha
Terra	mundo, planeta	continente, país, cidade	planeta do sistema solar	crosta, manto, núcleo	sistema solar

Analyzing Table 4, specifically for the word "cachorro," we can trace the following semantic relations and their terms: each term in the array of each relation has a connection with the word. For example, "pastor-alemão" is a hyponym of "cachorro"; "poodle" and "labrador" are also hyponyms of "cachorro." Similarly, "pata" and "cauda" are parts of "cachorro," making them meronyms of "cachorro." The same behavior is noted for "matilha" (a holonym of "cachorro").

However, when we look at the word "Terra" and its relations, this direction sometimes changes. For example, "continente" is not a hypernym of "Terra"; rather, "Terra" is a hypernym of "continente" or "país." On the other hand, "crosta," "manta," and "núcleo" are indeed meronyms of "Terra." This error appears in several instances of LLM-SEMREL. We believe it is due to prompt interpretation by LLM.

5. Conclusion

In this paper, we introduced LLM-SEMREL, a comprehensive semantic database for the Portuguese language, created using the capabilities of large language models (LLMs). This resource addresses the significant gap in semantic resources available for the Coreference Resolution task in Portuguese, providing a rich dataset of 1,229,399 semantic relations. By leveraging state-of-the-art LLMs, we have been able to generate a diverse and nuanced set of semantic relations that can enhance various NLP applications, including but not limited to coreference resolution, machine translation, and sentiment analysis.

Our error analysis revealed that while LLM-SEMREL is a robust resource, there are areas for improvement, particularly in the directionality of certain semantic relations. These errors highlight the challenges inherent in automatic annotation using LLMs and suggest avenues for refining our approach.

Looking forward, several directions for future work emerge. Firstly, refining the prompt design and interpretation mechanisms used by LLMs could mitigate the identified errors, enhancing the accuracy of the semantic relations. Additionally, integrating human-in-the-loop approaches for validation and correction could further improve the quality of the database.

Another promising direction is the expansion of LLM-SEMREL to include more semantic relations and a broader vocabulary. This could involve incorporating additional linguistic resources and leveraging advancements in LLMs to generate even richer datasets. Furthermore, applying LLM-SEMREL to other NLP tasks beyond coreference resolution could demonstrate its versatility and utility across different applications.

In conclusion, LLM-SEMREL represents a significant step forward in the creation of semantic resources for the Portuguese language. While there are challenges to address, the potential benefits for NLP applications are substantial, paving the way for more accurate and contextually aware language models. We look forward to the continued evolution of this resource and its impact on the field of NLP.

References

- Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.
- Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Yang, A., Fan, A., Goyal, A., Hartshorn, A., Yang, A., Mitra, A., Sravankumar, A., Korenev, A., Hinsvark, A., Rao, A., Zhang, A., Rodriguez, A., Gregerson, A., Spataru, A., Rozière, B., Biron, B., Tang, B., Chern, B., Caucheteux, C., Nayak, C., Bi, C., Marra, C., McConnell, C., Keller, C., Touret, C., Wu, C., Wong, C., Ferrer, C. C., Nikolaidis, C., Allonius, D., Song, D., Pintz, D., Livshits, D., Esiobu, D., Choudhary, D., Mahajan, D., Garcia-Olano, D., Perino, D., Hupkes, D., Lakomkin, E., AlBadawy, E., Lobanova, E., Dinan, E., Smith, E. M., Radenovic, F., Zhang, F., Synnaeve, G., Lee, G., Anderson, G. L., Nail, G., Mialon, G., Pang, G., Cucurell, G., Nguyen, H., Korevaar, H., Xu, H., Touvron, H., Zarov, I., Ibarra, I. A., Kloumann, I. M., Misra, I., Evtimov, I., Copet, J., Lee, J., Geffert, J., Vranes, J., Park, J., Mahadeokar, J., Shah, J., van der Linde, J., Billock, J., Hong, J., Lee, J., Fu, J., Chi, J., Huang, J., Liu, J., Wang, J., Yu, J., Bitton, J., Spisak, J., Park, J., Rocca, J., Johnstun, J., Saxe, J., Jia, J., Alwala, K. V., Upasani, K.,

- Plawiak, K., Li, K., Heafield, K., Stone, K., and et al. (2024). The llama 3 herd of models. *CoRR*, abs/2407.21783.
- Fonseca, E., Vieira, R., and Vanin, A. A. (2016). Improving coreference resolution with semantic knowledge. In *Computational Processing of the Portuguese Language - 12th International Conference, PROPOR 2016, Tomar, Portugal, July 13-15, 2016, Proceedings*, volume 9727 of *Lecture Notes in Computer Science*, pages 213–224. Springer.
- Fonseca, E. B. (2018). *Resolução de correferência nominal usando semântica em língua portuguesa*. PhD thesis. Escola Politécnica.
- Gonçalo Oliveira, H. (2012). *Onto. PT: Towards the Automatic Construction of a Lexical Ontology for Portuguese*. PhD thesis, Ph. D. thesis, University of Coimbra.
- Hedderich, M. A., Lange, L., Adel, H., Strötgen, J., and Klakow, D. (2021). A survey on recent approaches for natural language processing in low-resource scenarios. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2545–2568. Association for Computational Linguistics.
- Jiang, F. and Cohn, T. (2021). Incorporating syntax and semantics in coreference resolution with heterogeneous graph attention network. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1584–1591. Association for Computational Linguistics.
- Khosla, S. and Rose, C. (2020). Using type information to improve entity coreference resolution. In *Proceedings of the First Workshop on Computational Approaches to Discourse*, pages 20–31. Association for Computational Linguistics.
- Lima, T., Collovini, S., Leal, A. L., Fonseca, E., Han, X., Huang, S., and Vieira, R. (2018). Analysing semantic resources for coreference resolution. In *Computational Processing of the Portuguese Language - 13th International Conference, PROPOR 2018, Canela, Brazil, September 24-26, 2018, Proceedings*, volume 11122 of *Lecture Notes in Computer Science*, pages 284–293. Springer.
- OpenAI (2023). GPT-4 technical report. *CoRR*, abs/2303.08774.
- Reid, M., Savinov, N., Teplyashin, D., Lepikhin, D., Lillicrap, T. P., Alayrac, J., Soricut, R., Lazaridou, A., Firat, O., Schrittwieser, J., Antonoglou, I., Anil, R., Borgeaud, S., Dai, A. M., Millican, K., Dyer, E., Glaese, M., Sottiaux, T., Lee, B., Viola, F., Reynolds, M., Xu, Y., Molloy, J., Chen, J., Isard, M., Barham, P., Hennigan, T., McIlroy, R., Johnson, M., Schalkwyk, J., Collins, E., Rutherford, E., Moreira, E., Ayoub, K., Goel, M., Meyer, C., Thornton, G., Yang, Z., Michalewski, H., Abbas, Z., Schucher, N., Anand, A., Ives,

- R., Keeling, J., Lenc, K., Haykal, S., Shakeri, S., Shyam, P., Chowdhery, A., Ring, R., Spencer, S., Sezener, E., and et al. (2024). Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *CoRR*, abs/2403.05530.
- Speer, R. and Havasi, C. (2012). Representing general relational knowledge in conceptnet 5. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation*, pages 3679–3686.
- Ueda, R. (2005). Dicionário brispell. <https://github.com/fititnt/br.ispell-dicionario-portugues-brasileiro?tab=readme-ov-file>.

Automated Topic Annotation in Brazilian Product Reviews: A Case Study of Adversarial Examples with Sabia-3

Lucas Nildaimon dos Santos Silva¹, Livy Real²

¹ Department of Computing, Federal University of São Carlos, Brazil

² Federal University of Paraná, Brazil

lucas.silva@estudante.ufscar.br, livy.real@gmail.com

Abstract. *High-quality annotated data is essential for many Natural Language Processing tasks, but traditional human annotation methods are often resource-intensive. Large Language Models (LLMs) offer potential solutions by generating labels for training datasets. This paper explores the effectiveness of using the Sabia-3 LLM for automatically labeling data for a multi-label topic classification task in Brazilian Portuguese product reviews. We compare the performance of Sabia-3-generated labels against human annotations using the RePro dataset. The study evaluates Sabia-3 on both random and adversarial datasets, highlighting its strengths in frequent topics, while identifying limitations in more nuanced categories. Models trained on Sabia-3 annotations showed promising results in common categories but faced challenges with ambiguous cases. Our findings suggest that while LLMs can streamline parts of the annotation process, human oversight remains essential, particularly in complex or less frequent cases. This research contributes new insights into the use of LLMs for automated data annotation in Brazilian Portuguese.*

1. Introduction

High-quality annotated data has long been a critical challenge in Natural Language Processing (NLP). Traditional human annotation is both time-consuming and costly, often requiring specialized knowledge that can be difficult to obtain, particularly in niche fields such as legal, medical, or industrial sectors like oil and gas. These constraints make acquiring high-quality data challenging in both academic research and resource-constrained industries.

With the advent of Large Language Models (LLMs), many traditional NLP tasks are being reevaluated. LLMs are increasingly applied to end-to-end tasks, particularly those involving creativity or text generation, such as conversational agents. However, for many downstream tasks—like spam detection, document classification, and sentiment analysis—that already benefit from classical algorithms when trained on appropriate data, substituting LLMs may not be necessary or practical.

Instead, for many downstream tasks not related to generative applications, it is often more advantageous to use LLMs to generate data for training classical models rather than replacing established NLP pipelines with LLMs. Nevertheless, relying on LLM-generated labels poses risks, especially in the absence of a human-labeled test set for validation. The state-of-the-art performance of various models across different languages

and domains is not fully understood, and LLM capabilities are rapidly evolving, making current benchmarks potentially obsolete in the near future.

This paper contributes to this ongoing discussion by focusing on a specific application: multi-label topic classification in Product Reviews in Brazilian Portuguese. We compare human-annotated data with LLM-annotated data and assess the performance of models trained on these datasets. Our approach includes a method for balancing random samples with an adversarial dataset.

Product reviews, crucial in the e-commerce and marketplace sectors, significantly influence consumer purchasing decisions. Given the availability of review datasets in Brazilian Portuguese, our findings are not only relevant to this industry but also applicable to other contexts.

Our study focuses on evaluating the Sabiá-3 LLM [Almeida et al. 2024] for the annotation of a multi-label topic classification task in Brazilian Portuguese product reviews. We use the RePro dataset [dos Santos Silva et al. 2024] and compare the performance of Sabiá-3 with human annotators on both general and complex cases.

The key question driving this study is whether Sabiá-3, given the same guidelines as human annotators, can perform topic labeling with comparable accuracy and consistency. By comparing its performance against human annotations on both general and challenging adversarial datasets, we aim to provide insights into its viability as a replacement for human annotators in this specific task.

Our objectives were threefold: to compare human and LLM-generated annotations, to evaluate Sabiá-3's performance on both random and adversarial datasets, and to assess the quality of models trained on human-labeled and LLM-labeled data.

2. Related Works

We investigated two primary areas: prominent datasets of Product Reviews in Portuguese and recent advancements in using Large Language Models (LLMs) for data annotation.

Regarding Product Reviews, a well-established textual genre on the web [Pollach 2006], multiple datasets are available in Brazilian Portuguese. The earliest is the ‘Brazilian E-Commerce Public Dataset by Olist’¹, released in 2018. This dataset encompasses approximately 100,000 orders from 2016 to 2018, including details on order status, pricing, product attributes, and customer reviews. In 2019, [Real et al. 2019] introduced the B2W-Reviews-01 dataset, which contains over 130,000 product reviews and includes additional information such as reviewers’ gender, age, and location, along with product evaluations like a 5-star rating and a “recommend-to-a-friend” question answered by all reviewers.

Several studies have built upon B2W-Reviews-01. [Real et al. 2020] conducted the first analysis of topics within product reviews, while Brands.Br² [Fonseca et al. 2020] incorporated brand information to fill gaps in B2W-Reviews-01. [Zagatti et al. 2021] focused on anonymizing the B2W-Reviews-01 corpus to ensure compliance with the General Data Protection Law.

¹<https://www.kaggle.com/datasets/olistbr/brazilian-ecommerce>

²<https://github.com/metalmorphy/Brands.Br>

[dos Santos Silva et al. 2024] offers a comprehensive examination of topics in product reviews, extending the work of [Real et al. 2020]. It introduces the RePro corpus, a 10,000-sample subset of B2W-Reviews-01, annotated with topics by human experts. This corpus is available for non-commercial use on GitHub³ and HuggingFace⁴ under the CC BY-NC-SA 4.0 license.

Regarding LLMs for data annotation, [Ye et al. 2022] introduced ZEROGEN, enhancing zero-shot learning by generating task-specific datasets with LLMs for efficient inference. [Ding et al. 2023] evaluated GPT-3’s annotation performance, comparing it with traditional methods across various NLP tasks.

In contrast to previous works, our study focuses on evaluating the Sabiá-3 model for data annotation in Brazilian Portuguese, using a real-world dataset (RePro). While prior research explores LLMs like GPT-3 and GPT-4, often in English and for zero-shot or synthetic data generation, we provide a direct comparison between human and LLM annotations in a less-studied language. This contributes new insights into the effectiveness of LLMs for data annotation in non-English contexts.

3. Methodology

We selected Sabiá-3 for this study as it is the latest Brazilian LLM specifically trained on Brazilian Portuguese data. Its predecessor, Sabiá-2, was evaluated across 64 diverse exams, ranging from university entrance exams to professional certification and graduate-level tests, where it outperformed GPT-4 in 23 of the 64 assessments [Almeida et al. 2024]. Additionally, [Inacio and Oliveira 2024] demonstrated Sabiá-2’s effectiveness in humor generation, showing it to be on par with rule-based approaches for this task. By evaluating Sabiá-3, we aim to contribute to the growing body of research on Brazilian LLMs and further explore their capabilities.

To conduct our experiments, we selected two subsamples from RePro: a random sample representing general cases, and an adversarial dataset⁵ where human annotators disagreed on the assigned topics. For both samples, we re-annotated the reviews with Sabiá-3, following the original topic annotation guidelines. We then trained two models: one using human-labeled data and another using LLM-generated labels. Both models were evaluated against human-labeled test data to assess the quality of the annotations and the effectiveness of the trained models.

3.1. Dataset

We used the annotated samples from the RePro dataset, which consists of product reviews in Brazilian Portuguese. Each review in RePro is annotated with one or more of the following six topics:

- **ANÚNCIO (Advertisement):** Contexts where the delivered product aligns or misaligns with the information presented on the product’s webpage, such as description, images, technical specifications, and overall advertising.
- **PRODUTO (Product):** Comments on product quality, originality, value for money, attributes, user experience, and general compliments.

³<https://github.com/lucasnil/repro>

⁴<https://huggingface.co/datasets/lucasnil/repro>

⁵Available at: <https://github.com/lucasnil/repro>

- **ENTREGA (Delivery):** Related to the speed of delivery, time, non-delivery, in-store pick-up, virtual delivery (e.g., gift cards, codes), and freight comments.
- **CONDIÇÕES DE RECEBIMENTO (Receipt Conditions):** Comments about the state of the product upon receipt, such as damage, packaging quality, incomplete or incorrect orders, and whether the product met the customer’s expectations.
- **OUTROS (Others):** Contexts involving seller inquiries, customer service, stock availability, shopping experience, payment methods, or nonsensical information that is not harmful to the company.
- **INADEQUADA (Inadequate):** Harmful information, such as profanity, competitor mentions, legal references, external links, or personal information.

Text	Topic
Adorei, A caixa veio bem amassada mas o produto estava em ordem.. já usei e gostei	CONDICIONESDERECEBIMENTO, PRODUTO
<i>I loved it, the box came quite dented, but the product was in good condition.. I've already used it and liked it</i>	RECEIVINGCONDITIONS, PRODUCT
A cor desse celular não é dourado igual da imagem da foto, ele é beecem mais claro!!	ANUNCIO
<i>The color of this phone is not gold like in the picture, it's muuuch lighter!!</i>	ADVERTISEMENT

Table 1. Examples of RePro

Two distinct subsamples of the RePro dataset were used in the LLM annotation experiment:

- **Random Sample:** A random selection of 1400 reviews from the RePro dataset, providing a broad spectrum of product feedback.
- **Adversarial Sample:** A subset of 2454 reviews in which the original human annotators disagreed on the assigned topic labels, requiring a third annotator to resolve the conflicts. We hypothesize that this makes the dataset more challenging for automatic labeling.

3.2. LLM Annotation Prompt

The automatic annotation process was conducted using the Sabiá-3 LLM, which was tasked with assigning one or more of the six predefined topics to each review. The prompt used for the labeling task (translated from Portuguese) was as follows:

“You are an automatic product review labeler for an e-commerce platform. You must read and label reviews with one or more of the following six topics: ANÚNCIO, PRODUTO, ENTREGA, CONDIÇÕES DE RECEBIMENTO, OUTROS, and INADEQUADA. To do this, you must strictly follow the annotation guidelines provided.”

The annotation guidelines, included with the prompt message, correspond to the topic descriptions outlined in Section 3.1. Each review was presented to the model via this prompt, and the Sabiá-3 LLM generated responses. The Sabiá-3 model was integrated through the Langchain framework, utilizing the Sábria-3 API to carry out the labeling task.

3.3. Human vs. LLM-Generated Labels

To assess the impact of label quality, we fine-tuned two BERT models for multilabel classification: one using human-annotated labels and another using labels generated by the Sabiá-3 LLM. Both models were trained on the Random Sample dataset and evaluated on a separate test set of 1000 human-labeled samples.

We employed BERTimbau, a pre-trained BERT model for Brazilian Portuguese. The models were fine-tuned using standard hyperparameters optimized for multilabel classification. Training was conducted for 3 epochs with a learning rate of $2e-5$, a batch size of 16 (training) or 64 (evaluation), and a weight decay of 0.01. Model performance was monitored every 10 steps, with the best-performing model (based on F1 score) saved to ensure optimal balance between precision and recall across labels.

4. Results

In this section, we present the results achieved by Sabiá-3 on the task of automatically labeling product reviews with multilabel topics and the comparison of models trained on human and LLM-generated labels.

4.1. Automatic Labeling

The results for the Random Sample dataset and the Adversarial dataset are displayed in Table 2 and 3, respectively. We evaluate the LLM using accuracy, precision, and F1-score for each class.

Class	Precision	Recall	F1-Score	Support
ANUNCIO	0.62	0.83	0.71	121
ENTREGA	0.93	0.98	0.96	455
PRODUTO	0.94	0.96	0.95	1087
CONDICOES DE RECEBIMENTO	0.55	0.80	0.65	229
INADEQUADA	0.25	0.45	0.32	58
OUTROS	0.86	0.41	0.56	344
Macro average	0.69	0.74	0.69	2294

Table 2. Performance metrics for the Random Sample dataset

Class	Precision	Recall	F1-Score	Support
ANUNCIO	0.65	0.84	0.73	429
ENTREGA	0.87	0.96	0.91	947
PRODUTO	0.82	0.90	0.86	1566
CONDICOES DE RECEBIMENTO	0.69	0.74	0.71	822
INADEQUADA	0.34	0.44	0.38	231
OUTROS	0.87	0.36	0.51	1118
Macro avg	0.71	0.71	0.68	5113

Table 3. Performance metrics for the Adversarial dataset

Across both datasets, the Sabiá-3 model demonstrated strong performance on the ENTREGA and PRODUTO classes, which consistently achieved high precision and recall values. These results highlight the model’s effectiveness in identifying frequent and well-defined topics. However, the model struggled significantly with the INADEQUADA and OUTROS classes, where both precision and recall were notably lower than for other classes. For INADEQUADA, the model exhibited very low precision, indicating a high rate of false positives, meaning that it frequently mislabeled reviews as inappropriate

when they were not. This could suggest that the model is overly sensitive to certain keywords or language patterns that it associates with harmful content, even in cases where human annotators would not. The recall for this class was also low, reflecting the model’s difficulty in identifying a substantial portion of truly inappropriate reviews. Similarly, for OUTROS, while the precision was relatively high, the recall was much lower, which implies that the model was conservative in assigning this label. It only identified a subset of the true instances of this class but was generally accurate when it did. This suggests the model may be underrepresenting miscellaneous topics that don’t fit cleanly into the other predefined categories.

4.2. Comparison of Models Trained on Human vs. LLM-Generated Labels

The results for both models, with human and LLM labels, are displayed in Tables 4 and 5, respectively.

Class	Precision	Recall	F1-Score	Support
ANUNCIO	0.97	0.51	0.67	72
ENTREGA	0.96	0.97	0.97	317
PRODUTO	0.96	0.94	0.95	774
CONDICOES DE RECEBIMENTO	0.91	0.69	0.79	180
INADEQUADA	0.00	0.00	0.00	34
OUTROS	0.85	0.64	0.73	235
Macro avg	0.78	0.63	0.68	1612

Table 4. Classification performance metrics for the model tuned with human labels

Class	Precision	Recall	F1-Score	Support
ANUNCIO	0.67	0.78	0.72	72
ENTREGA	0.95	0.97	0.96	317
PRODUTO	0.93	0.95	0.94	774
CONDICOES DE RECEBIMENTO	0.63	0.69	0.66	180
INADEQUADA	1.00	0.03	0.06	34
OUTROS	0.91	0.22	0.35	235
Macro avg	0.85	0.61	0.62	1612

Table 5. Classification performance metrics for the model trained with LLM labels

The model trained on human labels exhibited more balanced performance across most classes, particularly excelling in recall, indicating it identified more relevant instances. However, it struggled with the INADEQUADA class, which had the lowest metrics possible. In this case, this should be due to fewer training examples of this particular class. However, previous work, such as in [dos Santos Silva et al. 2024], showed that class INADEQUADA continued to perform poorly despite a larger number of samples, suggesting inherent challenges in labeling this class accurately, likely due to its ambiguity.

The model trained on LLM-generated labels demonstrated greater variability across classes. It maintained strong performance in common categories like PRODUTO and ENTREGA, but its precision for the OUTROS class was notably lower compared to

the human-labeled model, suggesting difficulty in distinguishing this category. In contrast, the LLM-generated labels yielded higher recall for the ANUNCIO class, indicating the model identified more instances but at the cost of precision, likely including more irrelevant cases. As with the human-labeled model, performance for the INADEQUADA class remained low.

5. Automatic Labeling Qualitative Analysis

In analyzing the quality of the Sabia-3 outputs, the most noticeable issue was the occurrence of hallucinations—69 cases out of 3,854 instances—and how these hallucinations manifested. We define hallucinations in two ways: when the model introduces a new topic not covered in the guidelines, and when it provides an explanation for a label.

Although we did not encounter any output that was entirely misaligned with the task, Sabia-3 often attempted to be more specific than necessary.

Consider the following example: **O produto é bom nos primeiros 6 meses, depois começa a dar problemas. Esse é o segundo que compro, pois o primeiro tive o dinheiro ressarcido, aí comprei este. Recomendo comprar sempre com a garantia estendida. Guardem a nota fiscal do produto e da garantia estendida. É funcional pois serve também para vigiar a casa e os pets.**⁶.

A human labeled this as: **OUTROS, PRODUTO**, since the review discusses the product, and warranty was explicitly categorized under the **OUTROS** topic in the guidelines. Sabia-3, however, produced the following output: **PRODUTO, GARANTIA (dentro de OUTROS), CONDIÇÕES DE RECEBIMENTO (referente à garantia e ressarcimento)**. While the model correctly labeled **PRODUTO**, it tried to be more specific by introducing a new label, **GARANTIA** (warranty), though it accurately recognized this as part of the **OUTROS** topic. The confusion arose with **CONDIÇÕES DE RECEBIMENTO**, which is a label intended to describe the state of the product upon receipt, not conditions after the product has been received.

Sabia-3 also generated more specific, albeit incorrect, labels such as **CUSTO-BENEFÍCIO, ATENDIMENTO, ESTOQUE, EXPERIÊNCIA DE COMPRA, ATENDIMENTO AO CONSUMIDOR**.⁷ Interestingly, all of these topics are closely related to the review content and relevant to the domain, suggesting these are domain-related hallucinations rather than out-of-scope hallucinations. However, despite their relevance, these labels were incorrect for the task at hand and would require careful post-processing. It's worth noting that domain-related hallucinations are significantly harder to detect than those that are completely unrelated to the task.

Lastly, consider the review: **EXCELENTE COMPRA Comprei esse motor e adaptei na vassoura elétrica. A patroa adorou (e eu também). Ligou o bicho, saiu voando por aí e até hoje não voltou.**⁸.

⁶*The product is good for the first 6 months, but then it starts having issues. This is the second one I've bought because I got a refund for the first one, so I purchased this one. I recommend always buying it with an extended warranty. Keep the receipt for both the product and the extended warranty. It's functional as it can also be used to monitor the house and pets.*

⁷*COST-BENEFIT, SUPPORT, STOCK, PURCHASE EXPERIENCE, CUSTOMER SERVICE*

⁸*EXCELLENT PURCHASE I bought this motor and adapted it to an electric broom. The wife loved it (and so did I). She turned it on, took off flying, and hasn't come back since.*

In this case, the model misclassified the topics but subtly acknowledged the humorous tone of the review, labeling it as *OUTROS (devido ao tom jocoso e à experiência do usuário)*⁹, indicating an attempt to account for the review’s humorous content.

6. Conclusions

In this study, we evaluated the robustness of Sabiá-3, a Brazilian Large Language Model (LLM), in the task of multi-label topic annotation for product reviews. By comparing its performance against human-labeled data, we investigated both general cases from a random sample and more challenging cases from an adversarial sample in the RePro dataset. Additionally, we trained models on both human and LLM-generated labels to assess the impact of label quality on downstream model performance.

Our findings indicate that Sabiá-3 performs well in identifying clear and frequent topics, such as ENTREGA and PRODUTO, but faces significant challenges in more ambiguous categories like INADEQUADA and OUTROS. These discrepancies underscore the limitations of LLMs when tasked with handling nuanced or infrequent cases, which require a more sophisticated understanding of context.

From the qualitative analysis, we observed that Sabia-3 occasionally produced hallucinations, introducing labels not covered in the original guidelines or over-specifying topics. While these hallucinations were often domain-relevant, they deviated from the task’s specific requirements, indicating the model’s tendency to infer context too aggressively. This over-specification led to issues in cases where the model introduced new labels like GARANTIA or misapplied existing ones, such as using CONDIÇÕES DE RECEBIMENTO inappropriately. These domain-related hallucinations are particularly concerning because they are harder to detect than out-of-scope errors, necessitating careful post-processing when relying on LLM outputs.

The comparison between models trained on human-labeled and LLM-labeled data further highlighted the challenges posed by ambiguous cases. While LLM-generated labels performed adequately in simpler categories, the human-labeled models provided a more balanced and accurate representation across all categories, particularly for complex or less frequent classes. This emphasizes the continued importance of human oversight in training data for high-stakes NLP tasks.

In conclusion, while LLMs like Sabiá-3 show promise in automating parts of the annotation process, particularly for well-defined and frequent topics, they struggle with edge cases and can introduce misleading labels. Our study contributes to the ongoing evaluation of LLMs, demonstrating the importance of including both general and adversarial datasets to test their limits. The hallucination issue also highlights the need for improved LLM interpretability and error correction mechanisms.

Future research should focus on developing advanced techniques for the automatic detection of adversarial examples within general datasets. Such methods could help streamline the annotation process by identifying challenging or ambiguous cases that typically require human intervention. This would not only reduce the reliance on human effort but also enhance the overall quality and reliability of LLM-generated annotations, enabling more efficient and accurate handling of complex tasks.

⁹*OTHERS (due to the playful tone and the user’s experience).*

References

- Almeida, T. S., Abonizio, H., Nogueira, R., and Pires, R. (2024). Sabiá-2: A new generation of portuguese large language models.
- Ding, B., Qin, C., Liu, L., Chia, Y. K., Li, B., Joty, S., and Bing, L. (2023). Is gpt-3 a good data annotator? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11173–11195.
- dos Santos Silva, L. N., Real, L., Zandavalle, A. C. B., Rodrigues, C. F. G., da Silva Gama, T., Souza, F. G., and Zaidan, P. D. S. (2024). RePro: a benchmark for opinion mining for Brazilian Portuguese. In Gamallo, P., Claro, D., Teixeira, A., Real, L., Garcia, M., Oliveira, H. G., and Amaro, R., editors, *Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 1*, pages 432–440, Santiago de Compostela, Galicia/Spain. Association for Computational Linguistics.
- Fonseca, E., Oliveira, A., Gadelha, C., and Guandaline, V. (2020). Brands.br - a portuguese reviews corpus. In *OpenCor*.
- Inacio, M. L. and Oliveira, H. G. (2024). Generation of punning riddles in portuguese with prompt chaining paper type: Late breaking results. *15th International Conference on Computational Creativity (ICCC'24)*.
- Pollach, I. (2006). Electronic word of mouth: A genre analysis of product reviews on consumer opinion web sites. In *Proceedings of the 39th Annual Hawaii International Conference on System Sciences (HICSS'06)*, volume 3, pages 51c–51c.
- Real, L., Bento, A., Soares, K., Oshiro, M., and Mafra, A. (2020). B2w-reviews02, an annotated review sample. In *OpenCor*.
- Real, L., Oshiro, M., and Mafra, A. (2019). B2w-reviews01-an open product reviews corpus. In *the Proceedings of the XII Symposium in Information and Human Language Technology*, pages 200–208.
- Ye, J., Gao, J., Li, Q., Xu, H., Feng, J., Wu, Z., Yu, T., and Kong, L. (2022). Zerogen: Efficient zero-shot learning via dataset generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11653–11669.
- Zagatti, F., Silva, L., and Real, L. (2021). Anonymization of the b2w-reviews01 corpus. In *OpenCor*.

Text extraction from Knowledge Graphs in the Oil and Gas Industry

Laura P. Navarro^{1,2}, Elvis A. de Souza^{2,3}, Marco A. C. Pacheco^{1,2}

¹ Department of Electrical Engineering (ELE/PUC-Rio)

² Applied Computational Intelligence Lab. (ICA/PUC-Rio)

³ Institute of Mathematical and Computer Sciences (ICMC/USP)

laurap@aluno.puc-rio.br, elvis.desouza99@gmail.com

Abstract. *This paper presents a detailed methodology for extracting and analyzing data from a knowledge graph designed to store complex geological information. Our pipeline was designed after a deep understanding of the KG, focuses on browsing, querying and transforming data using curated text templates. The extraction methodology is based on graph triples, key classes, properties and relationships, which ensures the relevance and truthfulness of the data obtained. With the recent advancements in neural large language models, which perform exceptionally well on open-domain tasks, our work addresses the challenge of presenting LLMs with accurate closed-domain data—originating from graph-based sources—in a readable and accessible textual format.*

1. Introduction

Efficient data extraction from knowledge graphs is fundamental for research and analysis across several fields [Ali et al. 2022]. Generating textual information from knowledge graphs involves identifying and extracting relevant information from the dataset, and may include collecting facts, relationships, and entities. Once this information has been extracted using SPARQL queries, it can be utilized to create coherent narratives or descriptions based on the extracted data [Ribeiro et al. 2020, Mizell et al. 2014, Koncel-Kedziorski et al. 2019].

Our goal is to generate accurate textual data to be ingested by a domain-specific Q&A system based on PetroKGraph, a knowledge graph from the oil and gas domain. Besides textual generation, we also generate questions and answers to evaluate the performance of our closed-domain Q&A system. In this context, the correct interpretation of data from a knowledge graph is crucial to understand the distribution and characteristics of geological formations and wells. Based on a small knowledge graph with 4,472 entities (and classes and relations) and using 29 text templates, we generate a dataset with 1,414 simple and 8,805 complex questions, along with 10,219 answers and respective contexts.

2. Related Work

Authors in [Peng et al. 2017] create a subgraph using keywords and then augment it using terms from SPARQL queries.¹ [Elbassuoni et al. 2010] search the knowledge graph

¹SPARQL [Harris et al. 2013] is a standard query language for finding and processing data in graph databases, it enables users to execute complicated queries to retrieve particular information. The main types of queries include SELECT, ASK and DESCRIBE to extract data by matching graph patterns.

dominated by extended SPARQL, which augments the triple patterns in SPARQL with keyword criteria. [Koncel-Kedziorski et al. 2019] explore and improve the process of generating multi-sentence text from automatically extracted information using knowledge graphs; in this work they introduce, GraphWriter, a model which leverages the structured representation of KG form to enhance the quality and coherence of generated texts.

Other approaches, such as graph-to-text generation, aim to produce fluent texts from graph-based data [Colas et al. 2022, Yuan and Färber 2023]. In order to assess the effects of various task-adaptive pretraining techniques in graph-to-text generation, [Ribeiro et al. 2020] examine pretrained language models, BART [Lewis 2019] and T5 [Raffel et al. 2020]. For our approach, we use text templates tailored by domain specialists, therefore the use of generative language models was not required.

3. PetroKGraph

Knowledge Graphs (KG) include various types of vertices and edges, which may be denoted as $G = (V, E, R)$. Here, V is a collection of vertices or entities, R is a collection of relationships, and each edge is specified by its relation type $r \in R$, which means that edges can be represented as triplets. RDF (Resource Description Framework) triples follow the structure of subject-predicate-object to represent data [Manola et al. 2004].

[Cordeiro et al. 2024b] present a methodology for extracting geoscientific entities and relations from technical documents, which was used to populate a knowledge graph called PetroKGraph, specifically developed for the oil and gas industry. The construction of PetroKGraph involved leveraging natural language processing (NLP) resources, including annotated corpora and embedding models, under the framework of Petro NLP [Cordeiro et al. 2024a]. This process was guided by the PetroKGraph Ontology, a geological ontology populated with classes, instances, subclasses, and relations extracted from technical documents and compiled by geoscientists. We used the PetroKGraph Ontology to generate the text dataset, and it used RDF to store the information.

4. Methodology

In order to extract contexts from PetroKGraph, a well-structured data extraction pipeline is required.² Figure 1 provides a systematic way to navigating, querying, and transforming data embedded in the RDF framework. The procedure begins with understanding the RDF graph structure, which is required to ensure that subsequent extraction methods are successful. By establishing the foundation with a thorough understanding of RDF principles, schema analysis, and graph exploration, we can efficiently construct a pipeline.

We start by understanding the RDF graph structure of PetroKGraph using Protégé software³ to visualize and explore its nodes, relationships, and entities. Also, we employ SPARQL to query and analyze the graph’s data patterns. We find that the PetroKGraph has 2,069 entities and 26 relations. Some of the entities and relations include: *basins*, *fields*, *well*, *texture*, *lithostratigraphic_unit*, *constituted_by*, *located_in*, *crosses*, *has_age*, *participates_in* and *part_of*.

To extract and analyze data from an RDF graph, we first set up a suitable environment by installing Python and essential libraries, specifically *rdflib*. After installation,

²We use *contexts* as textual statements encoding the knowledge from a KG in natural language.

³<https://protege.stanford.edu/>

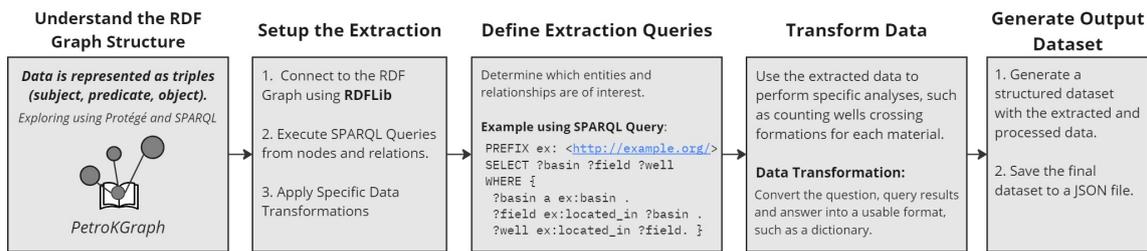


Figure 1. Pipeline for Data Extraction from RDF Knowledge Graph - PetroKGraph

we use *rdflib* to load the RDF graph into the environment. With a clear grasp of the RDF triples, key classes, properties, and relationships, we can now proceed to design and implement the pipeline. In this implementation phase, we determine which entities and relationships are of interest to effectively query, extract, and transform data. We write SPARQL queries and algorithms to extract relevant data from the PetroKGraph (Fig. 1).

We start transforming data by running SPARQL queries and algorithms as shown in next section 5. A dataset of context, question, and answer is created by these queries, which retrieve pertinent information based on criteria for Portuguese.

5. Experimental Results

Algorithm 1 is designed to process lithostratigraphic units or geological formation data, focusing on mapping materials to the formations they constitute. This is particularly useful in geological contexts, where the algorithm iterates through formations, extracts relevant material information, and automatically generates questions based on the relationships between geological formations and their materials.

Algorithm 1 Query to retrieve the materials that constitute each lithostratigraphic unit

```

1: materials_to_formation ← {}
2: for all formacoes, info in formations.items() do
3:   if info["constituted_by"] is not empty then
4:     formation_name ← info["labels"][0] if info["labels"] else formacoes
5:     for all material in info["constituted_by"] do
6:       material_id ← material.split("#")[1]
7:       material_name ← str(g.value(URIRef(namespace + material_id), rdfs.label))
8:       if material_name not in materials_to_formation then
9:         materials_to_formation[material_name] ← []
10:      end if
11:      materials_to_formation[material_name].append(formation_name)
12:    end for
13:  end if
14: end for

```

After processing all formations, the code iterates over the *materials_to_formation* dictionary as shown in Algorithm 2. For each material, it constructs a question about which lithostratigraphic units (formations) are constituted by that material. The question, along with the list of formations (as the answer) and a contextual statement, is stored in a questions list and saved in the dataset's JSON file.

Algorithm 2 Generate questions for lithostratigraphic units by material

```

1: for material_name, formation_name in materials_to_formation.items() do
2:   all_formation_names ← “,”.join(formation_name)
3:   formations_list ← [str(formacao) for formacao in formation_name]
4:   questions.append({
5:     “question”: f“Que unidades litoestratigráficas são constituídas por {material_name}?”,
6:     “answer”: formations_list,
7:     “context”: f“As unidades litoestratigráficas constituídas pelo {material_name} são:
   {all_formation_names}.”
8:   })
9: end for

```

Additionally, we used a different type of query, such as Algorithm 3, which retrieves information about basins from a SPARQL endpoint, basins (*?basin*) and lithostratigraphic units (*?lithostratigraphic_unit*) intersected by wells located in those basins. After processing all formations and basins, populates a list of question-answer pairs about the basins containing a specific geological formation.

Algorithm 3 SPARQL query to retrieve basins with specific lithostratigraphic units

```

1: SELECT ?basin ?litho_unit
2: WHERE {
3:   ?basin rdf:type ont:basin .
4:   ?litho_unit rdf:type ont:lithostratigraphic_unit .
5:   ?well ont:crosses ?litho_unit .
6:   ?well ont:located_in ?basin .
}

```

Table 1 presents a few text templates designed to extract specific information from PetroKGraph using the data extraction pipeline in Fig. 1. Each row represents a specific query formulated to obtain details about different aspects of the PetroKGraph, such as field locations and geological age of formations. This table illustrates the effectiveness of the queries and the applicability of the extraction method for collecting pertinent data.

Table 1. Text templates developed to extract information from PetroKGraph

id	Query	Relation
1	Onde está localizado o campo {campo_name}?	located_in
2	Onde atravessa o poço {well_name}?	crosses
3	Qual é a idade geológica de {formation_name}?	has_age
4	Qual é a entidade que faz parte de {formation_name}?	part_of

6. Conclusions

We develop a dedicated pipeline for extracting and analyzing complex geological data by utilizing the pre-existing PetroKGraph. We generated a dataset of 10,219 items, each consisting of a question, an answer, and a context. It will be used in a question answering (Q&A) system based on Retrieval-Augmented Generation (RAG), both for ingestion and for evaluating its performance in a closed-domain setting. This system will combine efficient data retrieval with contextual answer generation, enhancing the ability to provide accurate and relevant responses based on the information extracted from the PetroKGraph. Integrating this dataset into the Q&A system will enable smoother and more precise text generation from LLMs.

Acknowledgments

This work was supported in part by the Brazilian National Agency of Petroleum, Natural Gas and Biofuels (ANP) and the Brazilian National Council for Scientific and Technological Development (CNPq) under Grant 140506/2023-3.

References

- [Ali et al. 2022] Ali, W., Saleem, M., Yao, B., Hogan, A., and Ngomo, A.-C. N. (2022). A survey of rdf stores & sparql engines for querying knowledge graphs. *The VLDB Journal*, pages 1–26.
- [Colas et al. 2022] Colas, A., Alvandipour, M., and Wang, D. Z. (2022). Gap: A graph-aware language model framework for knowledge graph-to-text generation. *arXiv preprint arXiv:2204.06674*.
- [Cordeiro et al. 2024a] Cordeiro, F. C., da Silva, P. F., Tessarollo, A., Freitas, C., de Souza, E., Gomes, D. d. S. M., Souza, R. R., and Coelho, F. C. (2024a). Petro nlp: Resources for natural language processing and information extraction for the oil and gas industry. *Computers & Geosciences*, page 105714.
- [Cordeiro et al. 2024b] Cordeiro, F. C., Silva, P. F. d., Gomes, D. d. S. M., Souza, R. R., Coelho, F. C., and Ell, B. (2024b). Petro kgraph: A methodology for extracting knowledge graph from technical documents-an application in the oil and gas industry. *Available at SSRN 4776804*.
- [Elbassuoni et al. 2010] Elbassuoni, S., Ramanath, M., Schenkel, R., et al. (2010). Searching rdf graphs with sparql and keywords. *IEEE Data Eng. Bull.*, 33(1):16–24.
- [Harris et al. 2013] Harris, S., Seaborne, A., Prud’hommeaux, E., et al. (2013). Sparql 1.1 overview. *W3C recommendation*.
- [Koncel-Kedziorski et al. 2019] Koncel-Kedziorski, R., Bekal, D., Luan, Y., Lapata, M., and Hajishirzi, H. (2019). Text generation from knowledge graphs with graph transformers. *arXiv preprint arXiv:1904.02342*.
- [Lewis 2019] Lewis, M. (2019). Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.
- [Manola et al. 2004] Manola, F., Miller, E., McBride, B., et al. (2004). Rdf primer. *W3C recommendation*, 10(1-107):6.
- [Mizell et al. 2014] Mizell, D., Maschhoff, K. J., and Reinhardt, S. P. (2014). Extending sparql with graph functions. In *2014 IEEE International Conference on Big Data (Big Data)*, pages 46–53. IEEE.
- [Peng et al. 2017] Peng, P., Zou, L., and Qin, Z. (2017). Answering top-k query combined keywords and structural queries on rdf graphs. *Information Systems*, 67:19–35.
- [Raffel et al. 2020] Raffel, C., Shazeer, N., Roberts, A., Lee, K., et al. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

- [Ribeiro et al. 2020] Ribeiro, L. F., Schmitt, M., Schütze, H., and Gurevych, I. (2020). Investigating pretrained language models for graph-to-text generation. *arXiv preprint arXiv:2007.08426*.
- [Yuan and Färber 2023] Yuan, S. and Färber, M. (2023). Evaluating generative models for graph-to-text generation. *arXiv preprint arXiv:2307.14712*.

Getting Logic From LLMs

Annotating Natural Language Inference with Sabiá

Fabiana Avais¹, Marcos Carreira², Livy Real¹

¹Federal University of Paraná
Curitiba, PR – Brazil

²State University of Ponta Grossa
Ponta Grossa, PR – Brazil

avaisfabiana@gmail.com, marcscarreira@uepg.br, livyreal@gmail.com

Abstract. *We discuss the difficulties of annotation for Natural Language Inference in Portuguese, comparing human and Large Language Model annotations. We used 200 sentence pairs from the ASSIN2 dataset and re-annotated them for the inference task. A semanticist conducted the first annotation, and a second round was conducted using Sabiá-3, a large language model trained on Brazilian Portuguese data. We found that Sabiá-3 has the same agreement score as human annotators, but the LLM and human annotators disagree in cases involving different linguistic phenomena. While humans tend to disagree on pairs involving pragmatics or cultural knowledge, Sabiá-3 tends to mislabel sentences that share context but with no clear, logical relations among them. It shows that although LLMs are now statistically as effective as humans, LLMs and humans have different patterns for disagreement or mistaken annotations for Natural Language Inference.*

1. Introduction

Since the advent of classical machine learning methods, reliable annotated data has been the bottleneck of Natural Language Processing (NLP). Data augmentation techniques have been fruitful in developing and enhancing models for specific tasks, mainly when dealing with real-world scenarios.

Currently, the first wave of Large Language Models (LLMs) takes place. Considering the effort and time involved, LLMs are highly resource-intensive and expensive. In this scenario, using LLMs to label new data to train a classical model is feasible. However, evaluating the generated data is still an open challenge.

Here, we investigate the possibility of using Sabiá-3, a Brazilian LLM, for the Natural Language Inference (NLI) task. NLI is considered the initial step for semantic reasoning, and although it is elementary for humans, it remains a significant hurdle for machine reasoning.

Natural Language Inference is a task from Natural Language Understanding or Semantic Reasoning. It is, briefly, the task of inferring the validity of a given piece of text from another, i.e., ‘[s]uch inference can be defined as the process of concluding the truth of a textual statement based on (the truth of) another given piece of text’ [Dagan et al. 2013]. It is different from Classical Linguistics or Formal Semantics, in which inferences are seen as a strictly semantic phenomenon. It is also different from

implicatures, in which pragmatics and world knowledge are responsible for the relations derived between two informational pieces. Therefore, NLI is commonly simplified as the task of finding the relations between two sentences or two short texts, no matter which linguistic phenomena are involved in getting those relations [Condoravdi et al. 2003].

The first works on NLI date from the early 2000s with the Pascal Challenges [Dagan et al. 2006], when the task posed was Recognizing Textual Entailment (RTE). With the success of machine learning approaches that require large volumes of annotated data, larger datasets became available in the 2010s, such as SICK [Marelli et al. 2014], SNLI [Bowman et al. 2015] and MultiNLI [Williams et al. 2017]. In Brazilian Portuguese, we also have some work done on NLI. ASSIN shared task [Fonseca et al. 2016] and the SICK-BR dataset [Real et al. 2018] introduced the discussion on Portuguese, and some other works, such as ASSIN 2 [Real et al. 2020] and [Salvatore 2020] continued to discuss and advance the Brazilian Portuguese state-of-the-art. Nowadays, the theme is expected to be revisited with the advent of LLMs. [Bencke et al. 2024] showed that GPT-4 could classify simple entailment relations, highly agreeing with humans.

Thus, we intend to profoundly understand what LLMs can do for NLI in Portuguese. The primordial question we want to answer is how well LLMs, particularly Sabiá-3, can deal with Brazilian Portuguese inference. Since NLI is a subtle task that poses challenges even for humans, we also want to investigate how different the processing of corner cases for humans and LLMs is.

Finally, it is relevant to highlight that several works have pointed out inherent disagreements on NLI interpretation [Kalouli et al. 2017, Pavlick and Kwiatkowski 2019, Zhang and de Marneffe 2021, Kalouli et al. 2023]. It is also relevant to understand if this problem arises when using LLMs for NLI. This study also attempts to deeply understand inherent disagreements in Brazilian Portuguese and contribute positively to achieving noiseless NLI corpora in the future.

2. Related Works

In this section, we focus on works describing difficulties in NLI annotation, and also in NLI resources for Brazilian Portuguese.

[Kalouli et al. 2017] describes a manual investigation of the SICK Corpus [Marelli et al. 2014], reannotating part of the SICK data that were considered logically wrong as sentence pairs that were labeled as entailment in one direction and contradiction in another. This work pointed out the necessity of deeply understanding the linguistic phenomena involved in those cases. It led to [Kalouli et al. 2019], in which inference pairs were labeled by a group of graduate students on semantics that also provided explanations of their reasoning for each label. This work proposes a categorization of linguistics phenomena inherently difficult for humans to annotate, as *directionality* (a sentence can be inferred by another, but the opposite it is not true) and *loose definition* (a lexical item is polysemous, vague or ambiguous making it hard to ground the sentence meaning).

Finally, [Pavlick and Kwiatkowski 2019] discusses cases in which inherently human annotators would disagree, proposing that NLI systems should predict the distributions over human judgments and not categorical labels. [Zhang and de Marneffe 2021]

trained an Artificial Annotator to detect those inherent disagreements, simulating the uncertainty in the annotation process.

Considering the Brazilian Portuguese context, the ASSIN corpus [Fonseca et al. 2016] was the first dataset for NLI to appear. It was used in the *Avaliação de Similaridade Semântica e Inferência Textual*¹ shared task. ASSIN2 was the second edition of this challenge [Real et al. 2020]. Although both editions tested NLI and Semantic Similarity, we only focus on NLI here. The ASSIN2 dataset comprises 10k pairs of sentences that are human-labeled for inference and semantic similarity. It was thought to be as simple as possible: it has no named entities, and all the sentence pairs are in the present tense. Inference labels are simply ‘entailment’ or ‘none’, leaving aside other relevant labels for inference as ‘contradictions’. It makes it a good candidate for our investigation since this is the first attempt at using Sabiá-3 for NLI.

[Bencke et al. 2024] is a recent work describing the InferBR dataset, a NLI corpus created semi-automatically. It parts from previous datasets, as SICK-BR [Real et al. 2018], and it uses classical techniques to redesign the premises of the sentence pairs and GPT-4 to generate new hypotheses and label new pairs. The human evaluation shows that 99.9% of the assigned labels by GPT-4 are useful. Authors state that the ‘errors found are related to neutral boundaries with entailment and contradictions’ [Bencke et al. 2024, p.9056], which, we believe, are potentially related to inherent disagreements. These impressive results also inspired this work: we want to see if a Brazilian LLM would have similar results and which kind of pairs would be the most challenging.

Commonly, NLI is one of the tasks used to evaluate an LLM [Rodrigues et al. 2023, Chaves Rodrigues et al. 2023], but, to the best of our knowledge, there is no work digging into the applications of LLMs for NLI considering linguistic phenomena in Portuguese.

3. Methodology

We propose to evaluate how good the Sabiá-3 model, from the Sabiá family [Almeida et al. 2024], is for NLI. To do so, we took a sample of sentence pairs of ASSIN2, considered as our baseline. Then, two re-annotations were conducted, one by a semanticist and another by the Sabiá-3 model. Finally, we compare the results and discuss linguistic questions that arise.

Since LLMs are trained on vast amounts of public data and Sabiá-3 was not fine-tuned for the NLI task, our intention is not to describe the ‘LLM reasoning’ but to check the feasibility of using the model as one NLI annotator and describe the patterns we find in the annotated data.

Thus, our first step was blindly re-annotating the inference labels of 200 sentence pairs randomly selected from the ASSIN2. In the second step, we reannotated the same sample with LLM Sabiá-3 through MariTalk² platform. MariTalk is a free-of-charge chatbot that serves Sabiá models. Afterwards, we statistically analyzed the results of the three annotations, categorizing all disagreements by the semantic phenomena prominent in each

¹Evaluating Semantic Similarity and Textual Entailment.

²<https://chat.maritaca.ai/>, on August 28th, 2024.

sentence pair.

The hypothesis was that by re-annotating part of the corpus ASSIN2 we would find different inference labels in some pairs. Each pair of ASSIN2 was annotated from 3 to 5 people and the final labels were the ones with the majority agreement. One semanticist exclusively did our re-annotation to be grounded on logical relations and to leave aside pragmatical influence whenever possible.

Another investigation concerning the Sabiá-3 outputs took place, in which we compared its reannotation labels to ours and ASSIN2’s first annotation. Below there are some examples³ of ASSIN2 pairs and labels agreed upon by all three processes.

Premise	Hypothesis	Label
A senhora está mexendo ovos em uma tigela.	A mulher está mexendo ovos em uma tigela.	Entailment
Um homem está tocando violão.	Um homem está tocando o instrumento.	Entailment
Não tem água sendo bebida por um gato.	Um caminhão está descendo rapidamente um morro.	None
Não tem muitas pessoas no parque de patinação no gelo.	Muitas pessoas estão em um parque de patinação no gelo.	None

Table 1. Examples of ASSIN2.

For the Sabiá-3 classification, we created four different prompts and tested them on 10 pairs with different semantic phenomena. We tried out zero-, one- and few-shot prompting [Dang et al. 2022], using ASSIN2 samples as examples. Since the four prompts had the same results for the 10 analyzed pairs, we opted for the simplest one, the zero-shot prompt. The selected prompt was the following:

‘Você é um anotador da tarefa de acarretamento entre pares de sentenças. A relação de acarretamento acontece quando a partir de uma sentença [A] podemos concluir que uma outra sentença [B] também é verdadeira. Ou seja, de [A] podemos concluir [B]. Para cada par de sentenças, estabeleça a relação classificando-o como ENTAILMENT quando há acarretamento entre as sentenças e NONE quando não há acarretamento.’⁴

The prompt was also tested on both ChatGPT (GPT-4o mini) and Maritalk (Sabiá-3), classifying one, five, or ten pairs per request. Since we obtained the same results for these ten examples, we opted to go only with Sabiá-3 and to classify ten samples per request.

4. Results

Comparing the three outputs, the total number of sentence pairs with single disagreements, in which only one annotator disagrees with the other two, was 30 pairs out of 200. Within this set of differences, Sabiá-3 had 36.6% of the single disagreements; ASSIN2 had 13.3% of single disagreements; and finally, the semantic expert annotated 50% of the single disagreements.

³Translations of the examples: (A) The lady is stirring eggs in a bowl. (B) The woman is stirring eggs in a bowl. (A) A man is playing the guitar. (B) A man is playing the instrument. (A) There is no water being drunk by a cat. (B) A truck is going down a hill quickly. (A) There are not many people at the ice skating park. (B) Many people are at an ice skating park.

⁴‘You are an annotator of the entailment task between pairs of sentences. The entailment relation happens when from a sentence [A] we can conclude that another sentence [B] is also true. In other words, from [A] we can conclude [B]. For each pair of sentences, establish the relationship classifying it as ENTAILMENT when there is entailment between the sentences and NONE when there is no entailment.’

Considering inter-agreement annotations (ASSIN2 against LLM, ASSIN2 against semanticist, semanticist against LLM), the sum of mismatches within the groups was 60 disagreements. Sometimes, more than one annotator disagreed with the same pair, so the pair was listed in more than one group. In this scenario, 19 pairs formed the mismatches between the semanticist and the ASSIN2; 15 composed the disagreement between the LLM and the ASSIN2; and 26 pairs formed disagreement between the semanticist and the LLM.

Therefore, there were more annotation discrepancies between the semanticist and the ASSIN2, rather than the LLM and the ASSIN2. Consequently, it shows the LLM tends to agree more with the general labels from ASSIN2 than with the logically grounded analysis. This might show that the LLM captures a more general understanding of the task. However, [Davani et al. 2022] and [Uma et al. 2022] pointed out that aggregated labels, such as the ones considered golden in ASSIN2, often lead to an oversimplification of a given task, making the evaluation dataset less reliable.

The table above displaces group mismatches on annotations.

Group Disagreements	Quantity
ASSIN2 and Semanticist	19 pairs
LLM and Semanticist	26 pairs
LLM and ASSIN2	15 pairs
Sum	60 pairs

Table 2. Disagreement pairs

Considering ASSIN2 as a baseline, we can say Sabiá-3 performed the same as our specialist. The results also suggest the LLM is more aligned with the general understanding of the task seen in ASSIN2 labels than with a more logically grounded analysis. Since the disagreements did not occur in the same samples, we discuss these cases in depth in the next section.

5. Qualitative Analysis

Here we consider the linguistic aspects (such as semantics, pragmatics, lexical semantics, and syntax) more prominent on each pair with disagreement, considering the three annotations. Based on [Kalouli et al. 2019], we found the following categories: *loose definition*, *subevent*, *directionality*, *annotation error*, *interpretation of preposition*.

We take the label *loose definition* to categorize “concepts that are ‘loose’, subjective or vague to define” [Kalouli et al. 2017]. See one example from the subset below:

- (1) (A) Dois meninos no sofá estão jogando vídeo games.
 (B) Dois meninos estão no sofá jogando jogos na televisão.⁵

The example above is not a clear entailment (from A to B), yet there are relations between A and B. In most instances of playing video games, the event conceptually involves a screen exhibition, which can be thought of as a television. However, this is not

⁵(A)Two boys on the couch are playing videogames – (B)Two boys are on the couch playing games on the television.

a logical entailment, considering that nowadays video games can be played on other devices such as PCs, mobile phones, or handheld game consoles. Therefore, we understand that the boundaries between the definitions of ‘videogames’ and ‘games on the television’ are loose.

The category *subevent* was used to label pairs of sentences that could describe the same event, yet each one focuses on different moments, or subevents, of a wider event.

- (2) (A) Uma árvore está sendo apanhada por um homem.
(B) Um homem está carregando uma árvore.⁶

In this case, the verb in A, ‘apanhar’ (to pick), is considered part of the event ‘carregar’ (to carry), in the sense that the act of carrying only happens after picking up the object being carried, following [Parsons 1990]. This intuition may not align exactly with what a theory of event semantics would support, but it can be considered within the boundaries of semantics and pragmatics.

We categorized *directionality* for pairs in which B was more specific than A, therefore, by definition, one could not say that the pair had an entailment relation. It is well illustrated below:

- (3) (A) A mulher está tocando a flauta.
(B) Uma mulher está habilmente tocando uma flauta.⁷

In this case, ‘habilmente’ (skillfully) is a subset of the set in which women play the flute. For B being more specific than A, we do not consider an entailment from A to B, but there is an entailment from B to A.

We found some *Annotation errors*. This happens when information is probably misunderstood by the annotator. In the case of human annotators, it tends to happen in pairs with long lengths, in which only one element changes.

- (4) (A) Três meninos estão pulando nas folhas.
(B) Crianças em camisas vermelhas estão brincando nas folhas.⁸

It would be a clear mistake for the annotator to indicate that A entails B, as ‘camisas vermelhas’ (red shirts) in B adds a specific detail not present in A. While both sentences could describe children playing on the leaves, nothing in A implies or restricts the color of their shirts, as introduced in B.

We considered an *Interpretation of preposition* if the major difference between the pair’s sentences was due to a change in prepositions.

- (5) (A) Duas equipes estão jogando futebol **de** campo.
(B) Diferentes times estão jogando futebol **no** campo.⁹

⁶(A) A tree is being picked up by a man. – (B) A man is carrying a tree.

⁷(A) The woman is playing the flute. – (B) A woman is skillfully playing a flute.

⁸(A) Three boys are jumping on the leaves. – (B) Children in red shirts are playing on the leaves.

⁹(A) Two teams are playing soccer (football of field). – (B) Different teams are playing football on the field.

On one hand, in A the preposition ‘de’ (of) introduces a football type, a subset of the football sport played on a grass field (soccer). On the other hand, in B, the preposition ‘no’ (‘em’ + ‘o’ – ‘in the’) introduces the idea of place, but does not entail soccer.

From this categorization, the ‘loose definition’ category falls under the scope of the inherent disagreements [Pavlick and Kwiatkowski 2019]. The category ‘Interpretation of a preposition’ is hard to define. Logically, sentences of an entailed pair have different meanings, but in the context of NLI annotation tasks, these minor details (such as the interpretation of a single preposition) may be seen more as a prank than real data. It might have happened because part of the ASSIN2 corpus is a translation from the SICK corpus [Marelli et al. 2014] and some translations are just odd in Portuguese.

Most disagreements between the semanticist and the other annotations fall under the ‘loose definition’ category (10 of 15 cases). It suggests that ASSIN2 and Maritalk are more aligned with some pragmatic/contextual understanding of the language than a logician. An example of this is the pair:

- (6) (A) Um homem negro está andando perto de uma loja em uma cidade grande.
(B) Um homem negro está andando próximo a um prédio em uma grande cidade.¹⁰

This pair was labeled both by ASSIN2 and by Sabiá-3 as entailed, although a store is not necessarily a building.

Other disagreements between the semantic expert and other annotations fall under the ‘interpretation of preposition’ category, suggesting that these refined meanings were not covered by the LLM or the general Portuguese-speaking annotators. Example (5) was annotated as an entailment both by ASSIN2 and by Sabiá-3.

Concerning Sabiá-3 disagreements, we mostly found the ‘directionality’ category (6 out of 15), and the ‘subevent’ category (4 out of 15) taking place. First, we argue that there is no single pattern of disagreements on the LLM, but a plethora of phenomena. More specifically, the disagreements are not motivated by the same linguistic phenomena found in the logical grounded annotation. Considering the ‘directionality’ category, it may suggest that the model gets confused by the nature of the task, labeling a relation in a specific direction only and not considering the context of the pair. It also may explain the ‘subevent’ cases, since all the sentences in those cases are somehow related. The following examples were labeled ‘Entailment’ only by Sabiá-3.

- (7) (A) O cavalo está sendo montado por um homem.
(B) O homem está no passeio com o cavalo.¹¹
- (8) (A) Alguns homens estão jogando críquete.
(B) Um pequeno grupo de homens está alegremente jogando críquete.¹²

Finally, ASSIN2 labels were different from the two other annotations in cases of ‘loose definition’ (2 out of 4) and ‘annotation mistakes’ (2 out of 4).

¹⁰(A) A Black man is walking near a store in a big city. – (B) A Black man is walking near a building in a big city.

¹¹(A) The horse is being ridden by a man. – (B) The man is on a ride with the horse.

¹²(A) Some men are playing cricket. – (B) A small group of men is cheerfully playing cricket.

- (9) (A) Uma mulher está pegando uma lata.
(B) Uma mulher está agarrando uma lata.¹³
- (10) (A) Um homem está usando uma camisa azul e andando com os pés descalços em uma quadra de tênis.
(B) Uma pessoa está usando uma saia azul e andando descalça na quadra de tênis.¹⁴

Comparing the performance of Sabiá-3 and humans, Sabiá-3 got a human performance statistically. However, in the automated annotation, the mislabeled cases are not part of the ‘inherent disagreements’ cases that are expected to be problematic for humans. Sabiá-3 gets confused more often by the constraints of the task itself. It is also reasonable to say that, for the model, if the two sentences already shared enough contextual background, the model would find a relation between them. So, there is still room for Sabiá-3 models to get qualitative human performance.

6. Conclusions

To summarize, our study highlights annotation challenges in Natural Language Inference by comparing human and Large Language Model (LLM) annotations. Using 200 sentence pairs from the ASSIN2 dataset, we re-annotated the data with a semanticist and Sabiá-3, an LLM trained in Brazilian Portuguese. We then analyzed all the disagreements between the annotations and categorized them by the most prominent linguistic phenomenon involved in the pair.

While Sabiá-3 achieved the same agreement score as human annotators, the patterns of disagreement differed. Human annotators often diverged on pairs involving pragmatics or cultural knowledge, which are considered ‘inherent disagreements’. Sabiá-3 tended to mislabel pairs that shared some context but were not logical entailments. These findings suggest that although LLMs are now statistically comparable to human performance, they exhibit distinct patterns of error and disagreement, particularly in handling specific linguistic phenomena.

References

- Almeida, T. S., Abonizio, H., Nogueira, R., and Pires, R. (2024). Sabiá-2: A new generation of portuguese large language models.
- Bencke, L., Pereira, F. V., Santos, M. K., and Moreira, V. (2024). InferBR: A natural language inference dataset in Portuguese. In Calzolari, N., Kan, M.-Y., Hoste, V., Lenci, A., Sakti, S., and Xue, N., editors, *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.
- Bowman, S. R., Angeli, G., Potts, C., and Manning, C. D. (2015). A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*.
- Chaves Rodrigues, R., Tanti, M., and Agerri, R. (2023). Natural Portuguese Language Benchmark (Napolab).

¹³(A) A woman is picking up a can. – (B) A woman is grabbing a can.

¹⁴(A) A man is wearing a blue shirt and walking barefoot on a tennis court. – (B) A person is wearing a blue skirt and walking barefoot on the tennis court.

- Condoravdi, C., Crouch, D., De Paiva, V., Stolle, R., and Bobrow, D. (2003). Entailment, intensionality and text understanding. In *HLT-NAACL 2003 workshop on Text meaning*.
- Dagan, I., Glickman, O., and Magnini, B. (2006). The pascal recognising textual entailment challenge. In Quiñero-Candela, J., Dagan, I., Magnini, B., and d'Alché Buc, F., editors, *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment*, pages 177–190, Berlin, Heidelberg. Springer Berlin Heidelberg.
- Dagan, I., Roth, D., Sammons, M., and Zanzotto, F. (2013). Recognizing textual entailment: Models and applications. *Synthesis Lectures on Human Language Technologies*, 6(4):1–222. Publisher Copyright: © Morgan and Claypool Publishers. All rights reserved.
- Dang, H., Mecke, L., Lehmann, F., Goller, S., and Buschek, D. (2022). How to prompt? opportunities and challenges of zero- and few-shot learning for human-ai interaction in creative applications of generative models.
- Davani, A. M., Díaz, M., and Prabhakaran, V. (2022). Dealing with Disagreements: Looking Beyond the Majority Vote in Subjective Annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110.
- Fonseca, E., Borges dos Santos, L., Criscuolo, M., and Aluisio, S. (2016). Visao geral da avaliacao de similaridade semantica e inferencia textual. *Linguamatica*, 8(2).
- Kalouli, A., Real, L., and de Paiva, V. (2017). Textual inference: getting logic from humans. *Proceedings of the 12th International Conference on Computational Semantics (IWCS 2017)*.
- Kalouli, A.-L., Hu, H., Webb, A. F., Moss, L. S., and de Paiva, V. (2023). Curing the SICK and Other NLI Maladies. *Computational Linguistics*, 49(1):199–243.
- Kalouli, A.-L., Real, A. B. L., Palmer, M., and de Paiva, V. (2019). Explaining simple natural language inference. *Proceedings of the 13th Linguistic Annotation Workshop (LAW 2019)*, ACL.
- Marelli, M., Menini, S., Baroni, M., Bentivogli, L., Bernardi, R., and Zamparelli, R. (2014). A SICK cure for the evaluation of compositional distributional semantic models. In *Proceedings of LREC 2014*.
- Parsons, T. (1990). *Events in the semantics of English: A study in Subatomic Semantics*. MIT Press/Cambrige, London.
- Pavlick, E. and Kwiatkowski, T. (2019). Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics*, 7:677–694.
- Real, L., Fonseca, E., and Oliveira, H. G. (2020). Organizing the assin 2 shared task. *ASSIN 2 Shared Task: Evaluating Semantic Textual Similarity and Textual Entailment in Portuguese*.
- Real, L., Rodrigues, A., Vieira e Silva, A., Albiero, B., Thalenberg, B., Guide, B., Silva, C., de Oliveira Lima, G., Câmara, I. C. S., Stanojević, M., Souza, R., and de Paiva, V. (2018). Sick-br: A portuguese corpus for inference. In *Computational Processing of the Portuguese Language: 13th International Conference, PROPOR*

- 2018, Canela, Brazil, September 24–26, 2018, *Proceedings*, page 303–312, Berlin, Heidelberg. Springer-Verlag.
- Rodrigues, J., Gomes, L., Silva, J., Branco, A., Santos, R., Cardoso, H. L., and Osório, T. (2023). Advancing neural encoding of portuguese with transformer albertina pt-*. In Moniz, N., Vale, Z., Cascalho, J., Silva, C., and Sebastião, R., editors, *Progress in Artificial Intelligence*, pages 441–453, Cham. Springer Nature Switzerland.
- Salvatore, F. d. S. (2020). *Analyzing natural language inference from a rigorous point of view*. PhD thesis, Universidade de Sao Paulo.
- Uma, A. N., Fornaciari, T., Hovy, D., Paun, S., Plank, B., and Poesio, M. (2022). Learning from disagreement: A survey. *J. Artif. Int. Res.*, 72:1385–1470.
- Williams, A., Nangia, N., and Bowman, S. R. (2017). A broad-coverage challenge corpus for sentence understanding through inference. *arXiv*.
- Zhang, X. F. and de Marneffe, M.-C. (2021). Identifying inherent disagreement in natural language inference. In Toutanova, K., Rumshisky, A., Zettlemoyer, L., Hakkani-Tur, D., Beltagy, I., Bethard, S., Cotterell, R., Chakraborty, T., and Zhou, Y., editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4908–4915, Online. Association for Computational Linguistics.

Augmenting Data to Improve the Performance of Recommender Systems

Leticia Freire de Figueiredo^{1,2}, Joel Pinho Lucas¹, Aline Paes²

¹ Globo

Rio de Janeiro, RJ, Brazil

²Universidade Federal Fluminense

Niterói, RJ, Brazil

{leticia.figueiredo, joel.pinho}@g.globo, alinepaes@ic.uff.br

Abstract. News recommendation plays an essential role in suggesting news to users in a personalized way. Most commonly, Recommendation systems (RS) suggest new articles that closely align with topics and themes users have read or engaged with previously. In this context, RS typically benefits from news metadata, providing key news' attributes, enabling the system to find similar news in content, topics, and themes. However, in most production systems, metadata is often manually filled and may not accurately reflect the true context of the news. To address this challenge in video news, we propose an automatic annotation approach powered by BERTopic, enabling precise tagging of news content. The automatically annotated metadata is then applied within a collaborative filtering algorithm that leverages association rules, enhancing the system's ability to identify and recommend relevant news. The proposed approach was experimented within a case study in Globo, where it showed an improvement in video views for user sessions.

Resumo. A recomendação de notícias desempenha um papel crucial na gestão de notícias aos usuários de forma personalizada. A recomendação pode trazer notícias que sejam similares aos temas e tópicos presentes em artigos de notícias que o usuário leu no passado. Em geral, o processo de recomendação se vale de metadados anotados em cada notícia para descrever seus principais atributos. Entretanto, muitas vezes, estes metadados são preenchidos manualmente e podem não refletir de forma precisa o contexto da notícia. Para abordar este problema, propomos um processo de anotação automática para vídeos de notícias utilizando BERTopic. Os metadados anotados automáticos são, ao final, utilizados, em uma recomendação de filtragem colaborativa utilizando regras de associação. A abordagem apresentada foi testada em um caso de estudo na Globo, onde apresentou uma melhora na quantidade de visualizações de vídeos para sessões de usuário.

1. Introduction

News recommendations have the goal of suggesting news to users in a personalized way [Karimi et al. 2018]. For example, content-based recommender algorithms suggest items similar to those the user has engaged with in the past [Lops et al. 2011]. The pieces of news to be recommended - in our case, news articles or videos - are usually annotated

with metadata that describes their key attributes. However, when the documents' metadata is filled manually, they might not accurately reflect or entirely encompass the context of the document. Moreover, depending on the volume of historical news, manually annotating the metadata is laborious and error-prone, also potentially leading to a lack of accurate metadata.

Given the problem of incorrect tagging, this paper proposes annotating them automatically in news videos, replacing manual annotation to improve the accuracy and effectiveness of recommendations. To do this, we define metadata as news topics automatically extracted using BERTopic [Grootendorst 2022]. The proposed framework processes the title and subtitle of the video to generate topics that reflect the contextual content of the video. Then, the recommender algorithm leverages these annotations so that videos with similar topic distribution will most likely be of interest to the users.

We show the improvement in recommendation through a case study focused on Globo, the largest mass media group in Latin America. Globo's vertical information portals are responsible for providing informative content to more than 100 million unique daily users. In this context, recommendation engines are critical for enhancing user experience and driving publishing revenue based on articles and video consumption. This way, the challenge lies not only in devising recommendations for millions of users with varying engagement levels, profiles and content preferences but also in keeping recommendations relevant, as thousands of news articles and videos are published daily. Our results show a clear improvement in video views, when compared to recommendations based on manual annotation.

2. Case Study: BERTopic for augmenting recommendation metadata

BERTopic is a topic modeling technique that uses embeddings from a pre-trained language model to create clusters from a given *corpus*, while also maintaining the most important words in the topic descriptions for clearer interpretation [Grootendorst 2022]. In [Michiels et al. 2023], the authors used BERTopic to define a topic for each news article, to compute the variety of topics each user was exposed to, by counting the number of unique topics within an observation window.

For our case study, the *corpus* contains the titles and subtitles of each news video. After BERTopic generates the topics from it, each video receives its corresponding topic that will be later used in the recommendation system. The recommendation system is built on a collaborative filtering algorithm that uses association rules, which we refer to as co-occurrence. The association rules focus on identifying patterns that predict the occurrence of one item based on the presence of other items within a transaction [Amatriain et al. 2010]. In our framework, the transaction corresponds to the total number of views within a given period.

In this context, each co-occurrence is restrained to consider videos that share the same topic generated by BERTopic, instead of considering videos from the whole catalog. Within the collaborative filtering approach, the videos recommended to the user will be ones that share the same topic and co-occurred with previous videos they watched.

We compared our method in an AB experiment against an alternative using co-occurrence collaborative filtering, but using manually defined topics. The AB testing

was built with the following setup: 1) the control: a naive approach, recommending the most recent videos from the catalog; 2) the baseline alternative: the previous collaborative filtering approach using manual annotations; 3) the collaborative filtering algorithm implementation using BERTopic. We measure the lift between the alternative metrics with the control alternative, as shown in Table 1. The **Video views per session** metric represents, on average, how many video views there are per user in a session. In the **Conversion rate** metric, we calculate the Click Through Rate (CTR). This metric calculates how many clicks the recommendation obtained [Jannach and Jugovac 2019]. In Globo’s context, the CTR is used mostly as a guardrail metric, whereas the video views per session is the primary metric targeting user engagement.

The results demonstrate an improvement on the recommendation metrics when using automatic annotation, compared with the manual annotation. This indicates the recommendation became more accurate for the users.

Tabela 1. Lifts compared with the control alternative

	Alternative w/ automatic annotation	Alternative w/ manual annotation
Video views per session	7,04%	5,68%
Conversion rate	6,03%	5,34%

3. Conclusion

Recommendations in the news play an essential role in delivering personalized suggestions to users. Usually, a recommender algorithm uses document metadata to improve the recommendation. However, when the metadata is manually annotated, the recommendation will likely present bias and be irrelevant to the user because of incorrect tagging. In this paper, we proposed an automatic metadata annotation based on BERTopic, given a video news dataset. This metadata is used to improve an association rule collaborative filtering algorithm. The proposed solution was tested in a case study on Globo. The final results, after an AB experiment showed an improvement in click-through rate and video view metrics. As a future step, this approach will be experimented with annotating news articles, using different article parts - the article title, the article body, or both.

Referências

- Amatriain, X., Jaimes*, A., Oliver, N., and Pujol, J. M. (2010). Data mining methods for recommender systems. In *Recommender systems handbook*, pages 39–71. Springer.
- Grootendorst, M. (2022). Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.
- Jannach, D. and Jugovac, M. (2019). Measuring the business value of recommender systems. *ACM Transactions on Management Information Systems (TMIS)*, 10(4):1–23.
- Karimi, M., Jannach, D., and Jugovac, M. (2018). News recommender systems—survey and roads ahead. *Information Processing & Management*, 54(6):1203–1227.
- Lops, P., De Gemmis, M., and Semeraro, G. (2011). Content-based recommender systems: State of the art and trends. *Recommender systems handbook*, pages 73–105.

Michiels, L., Vannieuwenhuyze, J., Leysen, J., Verachtert, R., Smets, A., and Goethals, B. (2023). How should we measure filter bubbles? a regression model and evidence for online news. In *Proceedings of the 17th ACM Conference on Recommender Systems*, pages 640–651.

Brazilian Consumer Protection Code: a methodology for a dataset to Question-Answer (QA) Models

Aline Athaydes , Lucas Bulcao , Caio Sacramento , Babacar Mane ,
Daniela Barreiro Claro , Marlo Souza , Robespierre Pita

¹FORMAS Research Center on Data and Natural Language
Institute of Computing – Federal University of Bahia (UFBA)
Av. Milton Santos, s/n - Campus de Ondina – 40.170-110 – Salvador – BA – Brazil

{alineathaydes, lucasbulcao, caio.sacramento, babacarm}@ufba.br

{dclaro, msouzal, robespierre.pita}@ufba.br

Abstract. *This work introduces the methodology for building a new dataset based on the Brazilian Consumer Protection Code (CDC), focusing on question-answer (QA) models. The dataset collected legal data, including CDC articles, legal summaries, and court rulings from the Superior Court of Justice (STJ). Automated data extraction techniques using Python were employed, and advanced language models such as Llama3-8b-8192, Gemma2-9b-it, and GPT-4o-mini were used to generate question-answer (QA) structures. This work presents our methodology for creating such a dataset to be used by language models for training in the legal domain, particularly in the CDC domain.*

1. Introduction

The demand for judicial services in Brazil has grown enormously, especially with the high consumption rates in society. The customer-consumer relationship has become increasingly problematic and generated many legal issues, increasingly demanding the attention of legal professionals. Consequently, their practices are more costly, time-consuming, and even error-prone. One judicial challenge is facilitating the general public's access to legal information in easily accessible language. This holds particularly true for Consumer Protection Law (CPL). The large number of cases makes it unfeasible for the legal structure to provide a timely decision in their judgments.

The evolution of technologies over time is remarkable, with humans constantly seeking tools to facilitate work and optimize time. In the legal field, this reality is no different. The AI era has empowered how legal issues are addressed, particularly with Large Language Models (LLMs). Thus, new possibilities arise to optimize access to and interpret legal information.

Question and Answer (QA) models have been widely employed in the legal domain to provide legal tasks, such as simplifying questions, summaries, image captioning, etc. The Brazilian Consumer Protection Code (CDC, which in Portuguese means "Código de Defesa do Consumidor") is the main code of Consumer Protection laws in Brazil; despite being widely widespread, for instance, each commercial establishment must provide access to a copy for its customers, the CDC is still relatively unknown by the large population, which frequently raises questions from citizens and legal professionals regarding the rights of consumers and providers of services and goods. In this regard, a QA model on

CDC might simplify access to information, popularizing accessibility to the legal domain and simplifying legal comprehension by citizens. As demonstrated by the development of [Silveira et al. 2023], a model specifically designed for the Brazilian legal context, adapting language models to legal needs significantly improves accessibility to complex legal information, making it easier for non-experts to navigate within the law - a recent concern of Brazil's Federal Supreme Tribunal.

In the era of LLMs, a QA model must have access to vast data to provide valuable answers. Thus, our first step in advancing the legal domain of CPL in Brazil is building a dataset to act as the source for these answers, as discussed in [Rajpurkar et al. 2016], highlighting the importance of having structured data and the challenges faced in training QA models. The methodology of creating such a dataset is introduced in this work, as far as the evaluation is concerned, to acquire the quality, accuracy, and contextualization of the generated answers. This process involves a pipeline from collecting and organizing a large amount of data related to the CPL domain, including CDC's 118 articles, 40 Supreme Court's summaries, and 10,504 rulings related to Consumer Protection Law, and evaluating the data to provide a dataset with questions and answers that will be used to train a QA model in the future.

Within this work, we tried to answer the following Research Question:

- *How does the lack of specialized datasets focus on the Consumer Protection Code (CDC) impact the development of efficient Question-Answer (QA) models in the Brazilian legal domain?*
- *How do different language models (Llama, Gemma, and GPT) perform in generating questions and answers about the CDC, and what are the key performance differences between them?*

Our first main is to explore the creation of a structured dataset for QA in the domain of Consumer Protection Law, employing LLMs to create triples of questions, answers, and contexts based on the articles of the CDC, discussing challenges, data organization techniques, and impacts in the legal domain.

This article is structured as follows: Section 2 describes the related works on QA models applied to the legal domain, emphasizing the lack of datasets focused on the Consumer Protection Law. Section 3 provides the details of how we built the dataset, including data normalization, organization, and the process of generating questions and answers using language models in the GPT, LLaMA, and Gemma families. Section 4 compares the generated answers from the models and discusses the challenges and findings. Finally, the last section outlines the future directions of using this dataset to train QA systems for the legal domain, particularly within CPL-related issues.

2. Related Works

Developing pre-trained language models for legal domains is essential for creating efficient question-and-answer systems. The work of [Silveira et al. 2023] demonstrated the importance of adapting language models to the Brazilian legal context, utilizing various legal texts, such as the Federal Constitution and the Civil Code but it did not include the Consumer Protection Code (CDC). The absence of a dataset for the CDC emphasized the need to focus on CDC articles, legal summaries, and related court rulings to train and validate language models for protecting consumer rights.

Our nearest approach is carried by [Jardim et al. 2023], who creates a question answering system centered on the CDC. Inspired by the dataset’s methodology, we applied a similar strategy to generate questions and answers based on consumer protection law and jurisprudence. Their work showed that it is possible to structure a dataset from diversified textual sources, which was crucial in guiding the development of our dataset and focused on providing answers about consumer rights. Our approach is similar to [Jardim et al. 2023], where the authors created a question-answering dataset in the sports domain. The data extraction methodology and structuring it in a format suitable for question-answering systems served as a reference for our data organization.

Different from our approach, [Jardim et al. 2023] focuses on a sports-related domain rather than a legal one. While their methodology is valuable, they do not address the specific challenges and intricacies involved in legal contexts, particularly regarding the interpretation and application of legal documents like the CDC. Additionally, their work does not include court rulings or legal summaries, which are critical for providing a comprehensive understanding of consumer rights in Brazil.

From our knowledge, this is the initial dataset created for the CDC legal domain that has been included in a pipeline for Question-Answer (QA) tasks. Our work utilizes CDC articles, legal summaries, and court rulings to create a solid foundation for training language models that address consumer-related legal questions, filling a gap in the legal domain of consumer protection.

3. Method

Our data generation method is structured in a pipeline merging (1) consumer protection laws and jurisprudence, extracted from CDC, courts’ rulings and summaries, (2) a data normalization process, (3) QA examples generation with LLMs, (4) Prompt Adjustments and Results.

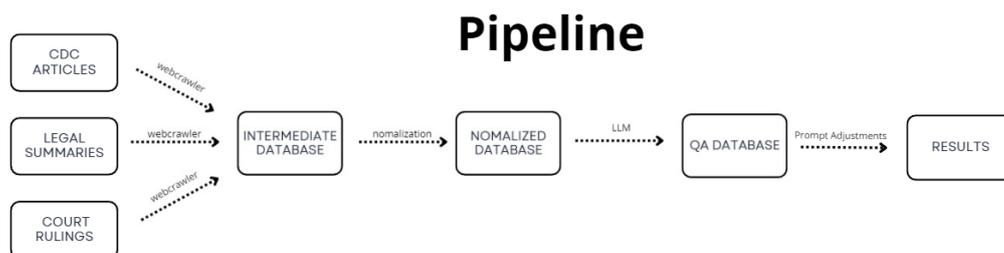


Figure 1. Methodology steps of our CDC dataset

3.1. Merging legal data of CDC

In the following, we discuss each step presented of the Pipeline depicted in Figure 1.

Firstly, we obtain the Brazilian Consumer Protection Code (CDC) directly from official websites through *web crawling* using *Selenium* with Python. The data is automatically structured on Titles, Chapters, Sections, and finally, Articles, based on the retrieved HTML structure. The complexity of the legal content was organized to make data manipulation easy for future steps.

Aiming to obtain a greater generality and applicability of our dataset, we expanded the initial set of laws extracted from the CDC with summaries and court rulings from the Superior Court of Justice (STJ). This expansion allows our dataset to better reflect the newest interpretation of consumer protection law and its applicability to different cases by incorporating relevant jurisprudence. We collected summaries and court rulings using the same web crawling approach, containing their respective numbers, descriptions, summaries, and detailed decisions. This process creates a more comprehensive and diversified dataset.

3.2. Data Normalization

After data collection, a textual normalization process was carried out to ensure consistency. This process is crucial to avoid issues during data manipulation, especially when performing automated analyses and feeding machine learning models. Managing accents and special characters in the Consumer Protection Code (CDC) legal texts, legal summaries, and rulings was one of the most challenging tasks we faced. These texts, often extracted from different sources, exhibited variations in encoding and formatting, which could compromise the quality of the final dataset.

We developed a Python function that utilizes the Unicode data library, known for efficiently handling strings. The function was designed to remove diacritical marks and special characters, transforming the text into a more standardized form and making it compatible with subsequent processing stages.

This normalization process is critical in legal contexts, where the accuracy of text representation can directly impact the interpretation of articles, summaries, and legal rulings. By standardizing the text and removing accents, we reduced inconsistencies.

The Superior Court of Justice (STJ) rulings represent a substantial part of the dataset and were particularly challenging due to their volume, with 10,504 rulings related to consumer protection. To facilitate processing and the generation of questions and answers, the rulings have been divided into 20 parts. We used the panda's library to handle these divisions and organize them into DataFrames.

3.3. Question Answer Generation with Language Models

To enrich our dataset, we used three language models for question-answer generation:

- **Gemma2-9b-it** [AI 2023]
- **Llama3-8b-8192**[Touvron et al. 2023]
- **GPT-4o-mini**[OpenAI 2023]

Adapted from [Jardim et al. 2023], we generated questions and answers based on the extracted CDC articles, summaries, and rulings. We focused on creating legally grounded questions and simulating scenarios. For this generation process, we connected our datasets to the models through their respective APIs and adjusted prompts, which coached the models to simulate common consumer doubts.

The development of [Silveira et al. 2023] directed our conceptual understanding to adapt pre-trained models to the Brazilian legal context. Unlike our approach, they do not include the CDC in their training, highlighting the importance of creating a dataset focused on the Consumer Protection Code, along with the STJ summaries and rulings.

3.4. Prompt Adjustments and Model Results

Adjusting the prompts was crucial in ensuring the quality of the generated questions and answers. To compare models, it was essential to create diverse questions and answers. For instance, there are three different text generation models: **Llama3-8b-8192**[Touvron et al. 2023], **GPT-4o-mini**[OpenAI 2023] and **Gemma2-9b-it** [AI 2023]. Every model received a carefully crafted prompt, with only minor alterations in keywords to reference CDC articles, legal summaries, or rulings.

The prompt adjustment was performed by requesting the generation of 10 to 30 questions for the Article of the CDC on the different models and evaluating the generated questions and answering, regarding informativeness and adequacy to the proposed use case, i.e. a user who is not a legal expert. Figure 2 shows the final prompt.

SYSTEM: Consider two distinct consumer groups: those who shop online and those who shop in physical stores. These consumers may have different levels of knowledge about the Consumer Protection Code (CDC), ranging from no knowledge to a basic or intermediate level, taking into account their shopping experiences and potential doubts. The questions should simulate real-life situations where they might need legal guidance or more details about how to ensure their rights. Focus on practical, everyday issues related to the CDC and situations these consumers might face. Never explicitly mention the SECTION in the questions or answers. The questions must be self-contained, meaning they should not require access to the section to be answered. Formulate questions technically, making them challenging and requiring a high level of understanding of the subject. Base the answers on LAWS, ARTICLES, RESOLUTIONS, and Rulings, explicitly integrating them into the ANSWER. For example: The Penal Code (art. X) stipulates the penalty...; Whenever there is mention of a law or article, integrate it into the ANSWER; Do not include personal opinions or speculations in the answers; Do NOT INCLUDE (Source:), (References:), (Based on section:), (Legal basis:) or similar. DO NOT CITE AUTHORS. DO NOT CITE THE SECTION.

USER: Assume the role of a Brazilian consumer with an issue involving online, and in-store purchases, or services provided by third-party companies, adopting the point of view of Brazilian consumers. Create a minimum of 10 and a maximum of 30 questions and answers based on " + {ementa} + ". Focus on real-life situations where they may need legal guidance, highlighting practical, everyday issues related to the Ruling. In addition to the questions and answers, bring the Ruling that supports the answer and its full description as context. Format the output exactly like this:

Question 1: [Text of question 1]
 Answer 1: [Text of answer 1]
 Context 1: [Description of Context ' 1]
 ...
 Question {n}: [Text of question {n}]
 Answer {n}: [Text of answer {n}]
 Context {n}: [Description of Context {n}]

Figure 2. Prompt employed for generation fo the triples question-answer-context in the dataset

By using this procedure, questions and answers were generated consistently and

then inserted into the corresponding data frame for each model.

4. Experiments and Results

In this section, we present our experiments and results with the three language models: **GPT-4o-mini**[OpenAI 2023], **Llama3-8b-8192**[Touvron et al. 2023] and **Gemma2-9b-it** [AI 2023]. We applied the same adjusted prompt to all models to ensure consistency in evaluating the results.

In this experiment, we provided CDC articles, legal summaries, and rulings as input to GPT-4o-mini, LLaMA3-8b-8192, and Gemma2-9b-it. Using an adjusted prompt, the models were fine-tuned to generate questions, answers, and corresponding contexts based on the legal data. We aimed to evaluate their ability to generate legally grounded responses in the consumer protection domain.

For each title, summary, and ruling, the models generated between 10 and 30 question-answer pairs. This allowed us to assess their performance in providing clear, accurate, and legally contextualized responses across various sections of the CDC.

4.1. GPT-4o-mini [OpenAI 2023]

We provide the following question to GPT-4o-mini:

Question : "What should I do if the product I bought online arrived defective?"

Answer : "You should contact the supplier immediately and request a replacement or repair of the product. If the defect is not resolved within 30 days, you can demand the substitution of the product, a refund of the amount paid, or a proportional reduction of the price."

Context : Article 18 of the CDC: "Suppliers of products and services are jointly liable for quality or quantity defects that render them unfit or inadequate for consumption..."

4.2. Llama3-8b-8192[Touvron et al. 2023]

We provide the following question to Llama3-8b-8192

Question : "What is my right of recourse in case of defective products or services?"

Answer : "According to Article 18 of the CDC, consumers have the right to receive products or services without defects. If the product or service is defective, the consumer can return to the company and request repair, replacement, or compensation, provided this is done reasonably and within the established deadline."

Context : The Consumer Protection Code aims to protect consumers' rights, ensuring they receive quality and safe products and services.

4.3. Gemma2-9b-it[AI 2023]

Question : "I bought a phone online, and upon opening the box, I noticed the product was damaged. The store claims the damage occurred during shipping. Do I have the right to request an exchange or refund even if the damage was not caused by the store?"

Answer : "Yes, It is within your rights to ask for an exchange or refund. Law No. 8.078/90, which establishes the Consumer Protection Code in Article 18, states that the product must be new, as advertised, and of quality. If the product arrived damaged, the store is responsible for providing it in perfect condition, regardless of the cause of the damage, as long as the consumer can prove the situation."

Context : Article 18 of the CDC: "The supplier is responsible for quality or quantity defects that render the products unfit or inadequate for consumption..."

4.4. Results and Discussions

In this section, we discuss the results of our experiments. To evaluate the models, we randomly selected 100 question-answer pairs from the dataset generated by each model and analyzed the proportion of incoherent, incomplete, or empty responses. These were classified into **adequate**, **regular**, and **irregular**. The dataset for this analysis is available in the **Data Analyze** repository, with code and dataset generation process documented ¹.

4.4.1. Model Comparison

GPT-4o-mini[OpenAI 2023]: produced fewer QA pairs compared to the other models, but the quality was generally higher. The majority of the responses were classified as **adequate**, with a small portion considered **regular**. The answers were clear, well-founded in the CDC, and followed the prompt instructions accurately. **Llama3-8b-8192**[Touvron et al. 2023]: generated over 50,000 questions and answers, but many of these were regular. The responses were often generic and repetitive, although the model successfully followed the prompt instructions. **Gemma2-9b-it** [AI 2023]: faced significant challenges, producing the highest proportion of irregular responses, including empty and incoherent answers. It struggled with formatting, factual accuracy, and adherence to prompt instructions, highlighting its limitations in handling legal data. In a targeted analysis requesting 10 irregular examples, only Gemma provided results, while GPT-4o-mini and Llama3-8b-8192 returned none, further confirming Gemma's deficiencies.

4.4.2. Dataset Unification and Conversion

After generating questions and answers for the different models, we unified all the resulting tables into a single **DataFrame**. This data frame contained more than 5,000 questions and answers, which were then converted into **JSON** format, widely used in natural language processing (NLP) model training due to its simplicity and flexibility.

4.4.3. Architecture and Fine-Tuning Tool

We used the **Axolotl** tool to fine-tune the model and prepare the dataset, which proved efficient for adapting language models to specific legal domains. The tool facilitated fine-tuning, allowing us to reduce training costs and time. We used the **JSON** data structure to ensure the models received well-organized information, improving the precision and efficiency of the fine-tuning process.

Similar to [Silveira et al. 2023], we adapted our models to handle the vocabulary and complexities of Brazilian legislation. While the focus of this work is the **dataset construction**, it provides a solid foundation for future implementations of **Questions-Answer (QA)** systems in the domain of the **Brazilian Consumer Protection Code (CDC)**.

¹<https://github.com/FORMAS/DIGGER>

5. Conclusions and Future Work

This work presented a methodology to create a dataset on the **Consumer Protection Code (CDC)**, incorporating articles, legal summaries, and court rulings. The complexity of the task required the use of various tools, including **Python**, **Selenium**, and advanced language models such as **Gemma2-9b-it** [AI 2023], **Llama3-8b-8192** [Touvron et al. 2023] and **GPT-4o-mini** [OpenAI 2023] to generate realistic and legally grounded questions and answers.

In future work, the goal is to use this dataset to train and fine-tune a model specifically designed for the **Question Answer (QA)** task related to the **CDC**. This model will have the ability to give precise responses to inquiries about consumer rights, facilitating both consumers and legal professionals with accessible legal information based on Brazilian law.

Acknowledgement

This work is partially financed by FAPESB CCE0022/2023, FAPESB TIC 002/2015, and CAPES Financial Code 001. We thank *Escavador.com* for their technical support.

References

- AI, G. (2023). Gemma2-9b-it model documentation. <https://gemma.ai>. Accessed: 2024-09-12.
- Jardim, P. C., Moraes, L. M. P., and Aguiar, C. D. d. A. (2023). Qasports: A question answering dataset about sports. *Repositório da Produção USP*.
- OpenAI (2023). Gpt-4 technical report. <https://openai.com/research/gpt-4>.
- Rajpurkar, P., Zhang, J., Lopyrev, K., and Liang, P. (2016). Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.
- Silveira, R., Ponte, C., Almeida, V., Pinheiro, V., and Furtado, V. (2023). Legalbert-pt: A pretrained language model for the brazilian portuguese legal domain. In *Proceedings of the Brazilian Conference on Intelligent Systems (BRACIS)*, pages 268–282. Sociedade Brasileira de Computação.
- Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., and Lample, G. (2023). Llama: Open and efficient foundation language models.